Generic Plurals Considered Harmful

Mark Liberman
University of Pennsylvania
https://www.ling.upenn.edu/~myl

ABSTRACT

This talk will sketch an interacting cluster of three fallacies:

1. Thinking of distributions as points;
2. Turning edges of distributions into taxonomic categories;
3. Forming stereotypes via socially-mediated confirmation bias.

These fallacies are all natural and nearly inevitable aspects of human thought.

But as a result, texts (scientific as well as popular) are full of statements about the properties and relationships of individuals and groups, which are somewhere between convenient shorthand (though sometimes for false or unrepresentative statistics) and total nonsense.

Some implications for "stochastic parrots" will be sketched.

Statistical language: the Pirahã and us

The Pirahã are a small tribe of hunter-gatherers in the Amazon basin.

Their language lacks not only the words but also the concepts for numbers.
Instead they use phrases with meanings like “small size”, “large size”, and “collection”.

The Pirahã people seem uninterested in learning about numbers,
and are even described as actively resistant to doing so.

…although in their dealings with traders,
they have a practical need to evaluate and compare numerical expressions.

Numbers are simple and natural concepts, of great practical importance.
How could rational people resist learning to understand and use them?

But we're not so different.

Until about 150 years ago, our language and culture lacked the words and ideas
needed to deal with the evaluation and comparison of sampled properties of groups.

And today, out of more than 300 million Americans,
maybe 50,000 to 500,000 have a solid grasp of the relevant statistical concepts and skills.

The rest of us are surprisingly uninterested in learning,
and actively resist the intermittent attempts to teach us,
although our frequent dealings with social and biomedical scientists
require us to evaluate and compare the numerical properties of representative samples.

A similar proportion of the 300-member Pirahã tribe
would yield 0.05 to 0.5 people who understand how to count and compare quantities
in the ways that have become important to their lives.

We often talk about sets as if they were individuals.

(…and never mind for now the problem of essentializing individuals…)

Some of these sets are what philosophers call “natural kinds”, like species of plants and animals,
but others are created by fiat,
out of combinations of circumstances and histories,
or regions of distributions,
or categories defined in even more arbitrary ways.

There are debates about how these sets should be defined and what they should be called.

But everybody talks about them as if they were individuals, at least sometimes,
because phrases like ”Xs are P-er than Ys” are just more convenient than phrases like

“The mean P measurement in a sample of Xs was greater than the mean P measurement in a sample of Ys, by an amount that would arise by chance fewer than once in 20 trials, assuming that the two samples were drawn from a single population in which P is normally distributed".

(…and generic statements also skip the often-tricky problem of defining the groups X and Y,
or making quantitative comparisons of more complex distributions...)

Caveat: comparisons of “generic plurals”
are not the only relevant types of generic statements.

We’ll mention several other potentially-misleading ways of summarizing complex data,
or of expressing stereotypes with or without a factual basis.

The limitations and complexities of ordinary language in this area
pose difficult problems for scientists, journalists, teachers, and everyone else.

But the problems are especially hard to avoid
for AI researchers aiming to turn large text collections
into an understanding of the world that the texts discuss.

To make it clearer what I'm talking about,
I'll give a few of the many concrete examples that I've studied in depth.

You can find more, every day, in the news or on social media.

These are complex examples,
and I'll present only a sketch of each one.
If you want the gory details for any of them,
ask me and I'll point you to a complete dissection.

I'll end by briefly discussing the semantics of generic statements in English,
and making some suggestions about ways forward for AI research.

Example 1 -- From a press release published on 6/1/2022:

Scientists Have Established a Key Biological Difference Between Psychopaths and Normal People

Neuroscientists using MRI scans discovered that psychopathic people have a 10% larger striatum, a cluster of neurons in the subcortical basal ganglia of the forebrain, than regular people. This represents a clear biological distinction between psychopaths and non-psychopathic people.

First, what are “psychopaths” and “normal people”?

According to the cited journal article,
“Psychopathy was assessed using the PCL-R, which consists of 20 items rated by interviewers on a 3-point *[0-2]* scale.”

The maximum score is 40 – and those scoring above 25 (or 30?) are classed as “psychopaths”.

The PCL-R items are diverse, ranging from “glibness”, “proneness to boredom”, and “promiscuous sexual behavior”, to “juvenile delinquency” and “revocation of conditional release”.

What about the associated brain regions?

According to the journal article, in MRI scans, “segmentation of the caudate, putamen, nucleus accumbens, and globus pallidus was conducted together with the thalamus and cerebellum using standard FreeSurfer parcellation. Total striatal volumes were defined as the sum of the volumes of the four striatal subregions.”

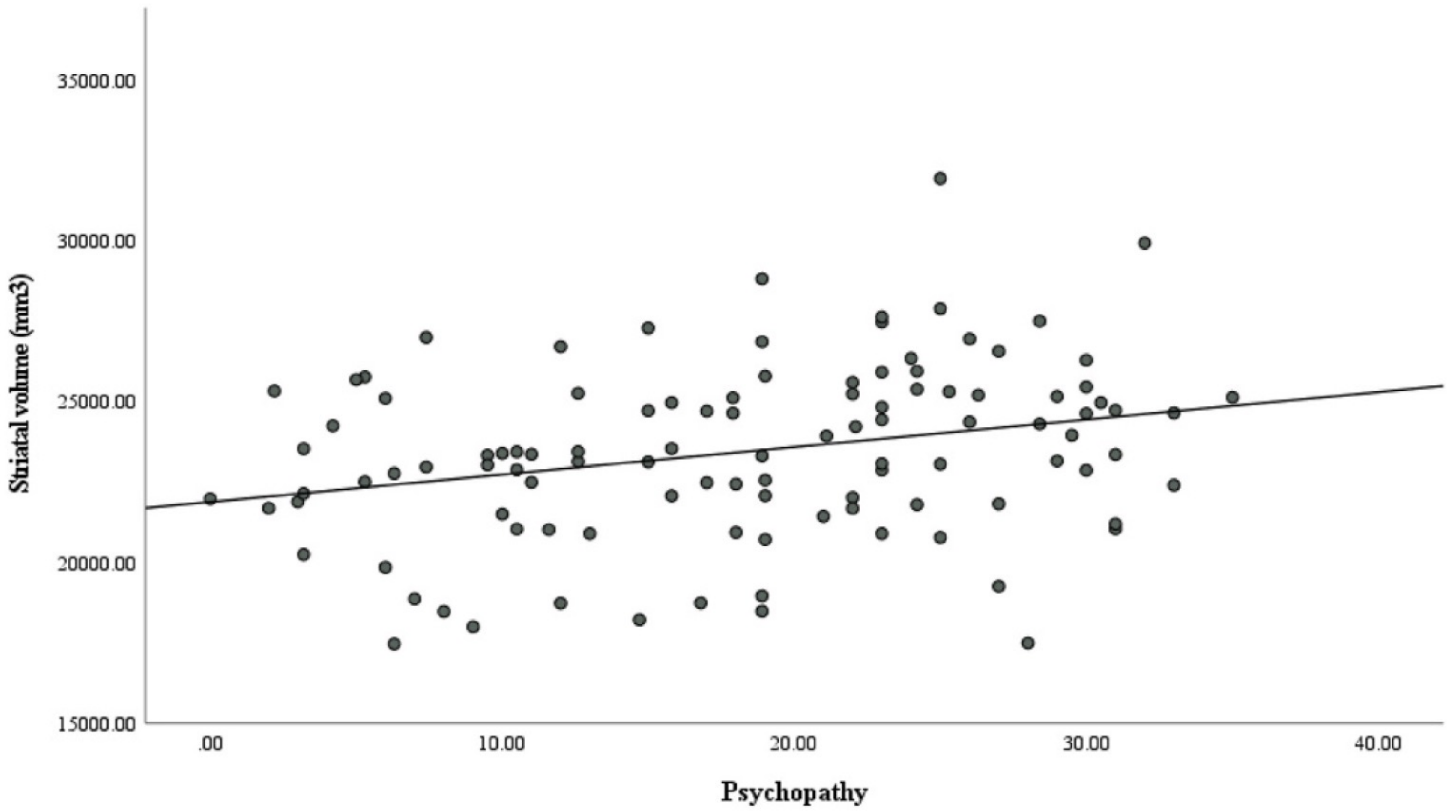
The generic plural "psychopaths" suggests a natural kind.

And the phrase "a clear biological distinction"
suggests well-defined and well-separated clusters of values
on both neuro-anatomical and social-psychological dimensions.

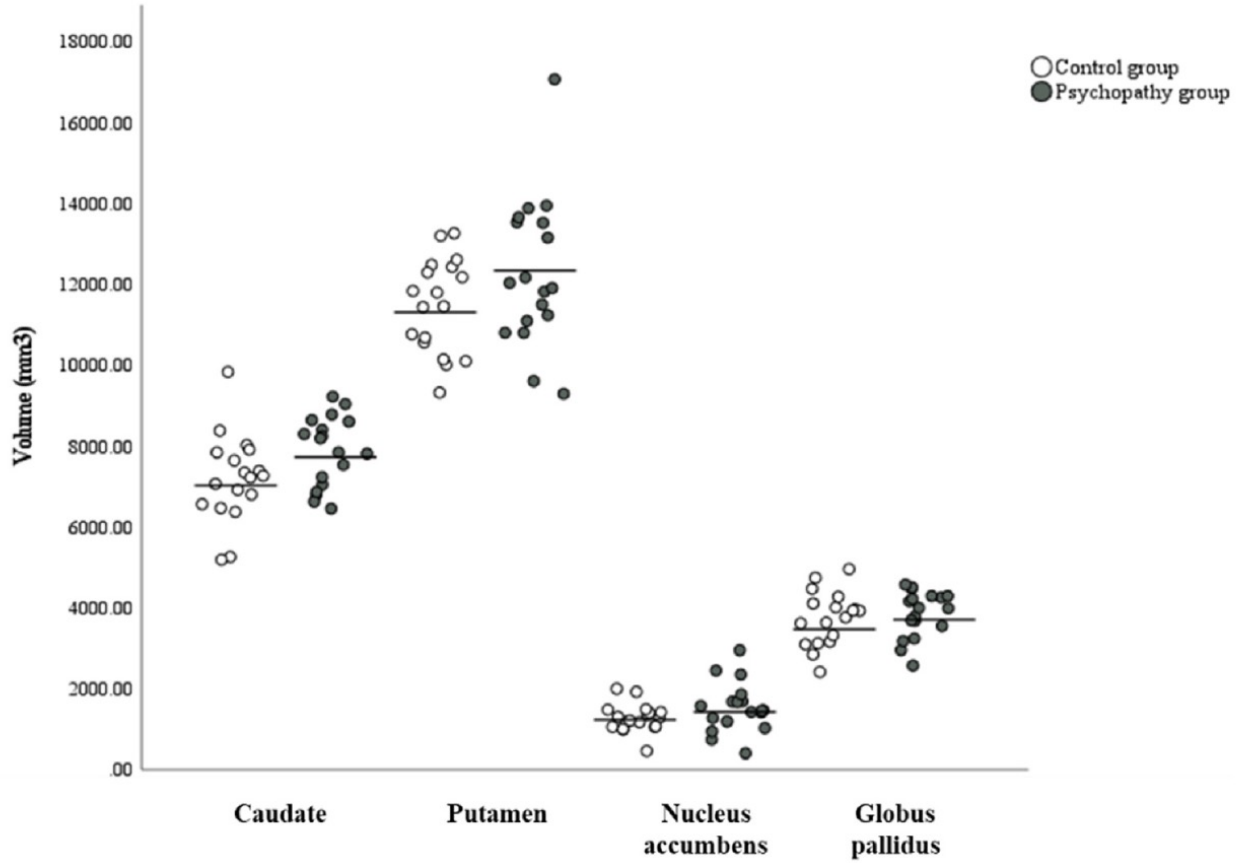
But what did the researchers actually find?

They found two weakly-correlated variables,
each an amalgam of several measurements or evaluations,
without any strong indication of clustering.

Their Figure 3,
showing all 108 subjects:



They selected 18 “psychopaths” and 18 controls from the ends of the PCL-R scale – and here’s what the brain volumes look like for those two sets:



This is all perfectly valid science –
at least it was accepted in a peer-reviewed journal.

But is *psychopath* really a “natural kind”?

Putting aside the obvious cases of racial, ethnic, gender, and other sociological stereotypes,
this way of thinking has been endemic in medicine for millennia.

It starts because boxing the world up into natural kinds is often a useful path to understanding.

And without convenient concepts and tools for dealing with continuous multi-variate distributions,
what else are you going to do?

Some medical conditions belong in well-defined categories.

But many behaviorally-defined disorders are "phenotypically diverse" —
and a clinician friend tells me that
"*phenotypically diverse* is the Greek translation of 'we have no fucking clue'".

"Phenotypic diversity" in this case:

The basic definition of *psychopathy* in the current official Diagnostic and Statistical Manual (DSM-5) is "three (or more)" of these seven characteristics:

1. Failure to conform to social norms with respect to lawful behaviors, as indicated by repeatedly performing acts that are grounds for arrest.
2. Deceitfulness, as indicated by repeated lying, use of aliases, or conning others for personal profit or pleasure.
3. Impulsivity or failure to plan ahead.
4. Irritability and aggressiveness, as indicated by repeated physical fights or assaults.
5. Reckless disregard for safety of self or others.
6. Consistent irresponsibility, as indicated by repeated failure to sustain consistent work behavior or honor financial obligations.
7. Lack of remorse, as indicated by being indifferent to or rationalizing having hurt, mistreated, or stolen from another.

There's some recent anti-boxology pushback from the profession --

"Objectives of the **Hierarchical Taxonomy of Psychopathology** (HiTOP) are to advance the classification of psychopathology to maximize its usefulness for research and clinical practice. The HiTOP aims to address limitations of traditional nosologies, such as the DSM-5 and ICD-10, including arbitrary boundaries between psychopathology and normality, often unclear boundaries between disorders, frequent disorder co-occurrence, heterogeneity within disorders, and diagnostic instability."

"The HiTOP approaches these problems by conducting an empirical search for psychopathology structures starting from the most basic building blocks and proceeding to the highest level of generality: combining individual signs and symptoms into homogeneous components or traits, assembling them into empirically-derived syndromes, and finally grouping them into psychopathology spectra (e.g., internalizing and externalizing)."

"The *[**NIH Research Domain Criteria**]* RDoC framework provides an organizational structure for research that considers mental health and psychopathology in the context of major domains of basic human neurobehavioral functioning, rather than within established diagnostic categories. The framework currently includes six major functional *domains.* Different aspects of each domain are represented by three to six psychological/biological dimensions, or *constructs*, which are studied along the full range of functioning from normal to abnormal."

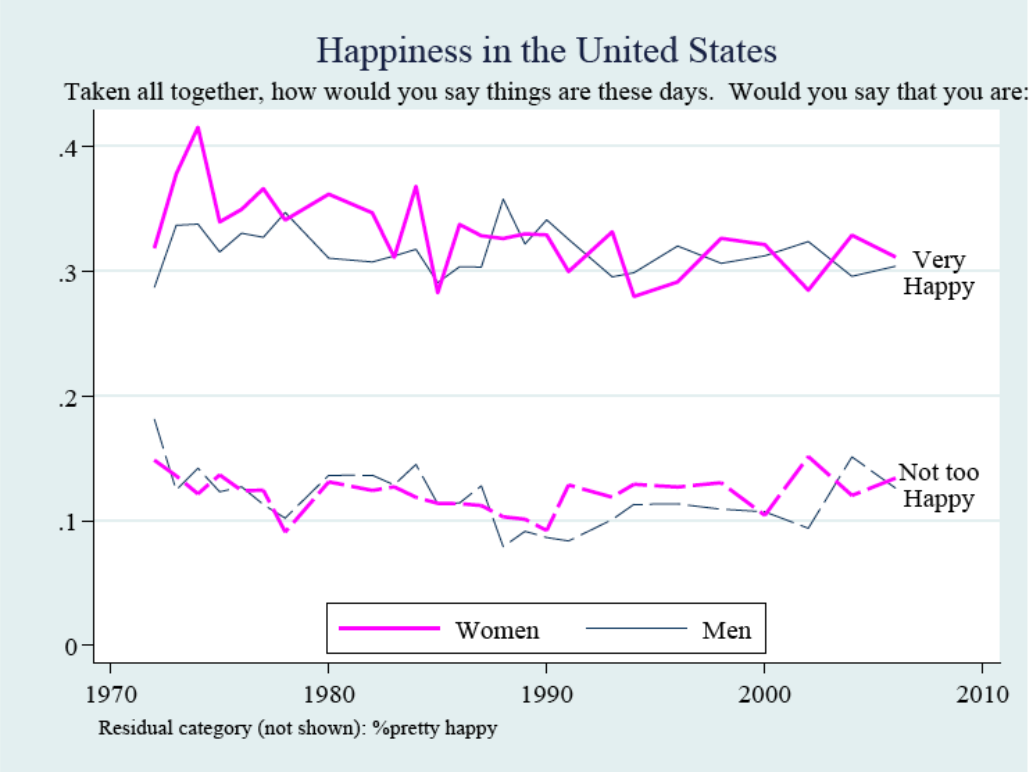
But it's hard to turn HiTOP or RDoC into clickbait – or even into an accessible press release.

Example 2 -- From a 2007 NYT article on "the growing happiness gap between men and women" :

"[E]conomists at the University of Pennsylvania […] have looked at the traditional happiness data, in which people are simply asked how satisfied they are with their overall lives. In the early 1970s, women reported being slightly happier than men. Today, the two have switched places."

The data --
from the *General Social Survey*:

Figure 1: Happiness in the United States, 1972-2006



How the economists themselves described the facts – a shift of 1.9 percentile points:

"Comparing the 2006 medians with the distribution for men in 1972, we see that the median woman in 2006 is as happy as a man at the 48.8th percentile in 1972 [...], while the median man in 2006 is as happy as the man at the 50.7th percentile in 1972."

But in their abstract, they wrote:

"By most objective measures the lives of women in the United States have improved over the past 35 years, yet we show that measures of subjective well-being indicate that women's happiness has declined both absolutely and relative to male happiness. [...] Our findings raise provocative questions about the contribution of the women's movement to women's welfare and about the legitimacy of using subjective well-being to assess broad social changes."

What people took away from the journalistic descriptions of this study
(and even from the journal article)
was that women used to be happier than men,
and now men are happier than women.

And they thought of this as a general fact about all men and all women.

But the effects were so small
that on an optimistic reading of a complex statistical reconstruction,
the happiest half of the population
contained maybe a couple of percent more of one sex than the other.

Leaving aside the problematic assumption of binarity,
gender-linked research often involves
similar inflation of small differences into important generic group properties.

About 15 years ago, there was boom in popular writing
on the magnitude and importance of ”sex differences”.

According to a popular book by Dr. Leonard Sax
(*Why Gender Matters: What Parents and Teachers Need to Know about the Emerging Science of Sex Differences*)

"Girls and boys play differently. They learn differently. They fight differently. They see the world differently. They hear differently."

And as a result, Sax argued, they need sex-segregated education.

Predictably, David Brooks in the NYT called this book “a lucid guide to male and female brain differences”.

Example 3 – focusing on the “hear differently” part of Sax’s theme.

Executive summary:

- He cherry-picked two unrepresentative numbers
  (out of 144 values in a 1959 table of hearing thresholds).
- He attributed to these numbers a subjective loudness difference of 100 to 1,
  whereas they actually represent a difference of 1.4 to 1.
- The chosen values compare young women and middle-aged men.
- He assumed that a higher hearing threshold (= hearing loss)
  means **better** speech comprehension in noise,
- and concluded that
  “Girls won’t learn as well in a loud, noisy classroom …
  [but] the rules are different when you’re teaching boys.”

Example 4 -- (alleged) ethnic differences in “individualism” vs. “collectivism”.

David Brooks in the *New York Times*:

“The world can be divided in many ways — rich and poor, democratic and authoritarian — but one of the most striking is the divide between the societies with an individualist mentality and the ones with a collectivist mentality.

This is a divide that goes deeper than economics into the way people perceive the world. If you show an American an image of a fish tank, the American will usually describe the biggest fish in the tank and what it is doing. If you ask a Chinese person to describe a fish tank, the Chinese will usually describe the context in which the fish swim.

These sorts of experiments have been done over and over again, and the results reveal the same underlying pattern. Americans usually see individuals; Chinese and other Asians see contexts.”

David Brooks is notoriously careless with facts,
but the relationship between this column and the cited research
reminded me of the old ”Ask Radio Yerevan” jokes from Stalinist Russia.

Question to Radio Yerevan: Is it correct that Grigori Grigorievich Grigoriev won a luxury car at the All-Union Championship in Moscow?

Answer: In principle, yes.

But first of all it was not Grigori Grigorievich Grigoriev, but Vassili Vassilievich Vassiliev;
second, it was not at the All-Union Championship in Moscow,
but at a Collective Farm Sports Festival in Smolensk;
third, it was not a car, but a bicycle;
and fourth he didn't win it, but rather it was stolen from him.

Question: Is it correct that if you show an American an image of a fish tank,
the American will usually describe the biggest fish in the tank and what it is doing,
while if you ask a Chinese person to describe a fish tank,
the Chinese will usually describe the context in which the fish swim?

Answer: In principle, yes.

But first of all, it wasn't a representative sample of Americans,
it was undergraduates in a psychology course at the University of Michigan;
and second, it wasn't Chinese, it was undergraduates in a psychology course
at Kyoto University in Japan;
and third, it wasn't a fish tank, it was 10 20-second animated vignettes
of underwater scenes;
and fourth, the Americans didn't mention the "focal fish" more often than the Japanese,
they mentioned them less often.

The research in question was reported in T. Masuda and R.E. Nisbett, "Attending holistically vs. analytically: Comparing the context sensitivity of Japanese and Americans", *J. Pers. Soc. Psychol.* 81:922–934, 2001.

The subjects were 36 Americans at the University of Michigan and 41 Japanese at Kyoto University, who "participated in the experiments as a course requirement".

The subjects each watched 10 animated vignettes of underwater scenes, where "Each scene was characterized by having 'focal fish,' which were large and had salient colors and shapes, moving in front of a complicated scene".

After a vignette was presented twice, the subject was asked "What did you see in the animation? Please describe it, taking as much as 2 min." The subjects' oral responses were recorded, transcribed and coded.

"The data were coded as belonging to one of the following categories: (a) focal fish, (b) background fish, (c) active animals, (d) inert animals, (e) plants, (f) bubbles, (g) floor of scene, (h) water, and (i) environment. … The categories were grouped into four superordinate categories. Focal fish remained an independent category. Background fish and active animals were grouped and named active objects, representing peripheral but moving objects. Inert animals and plants were categorized as inert objects. Finally, bubbles, floor of scene, water, and environment were categorized as background."

The results:

Table 2
Numbers of Accounts of Scenes in Each Category for American and Japanese Participants

	American			Japanese		
Category	*M*	*SD*	*n*	*M*	*SD*	*n*
Background	20.11	19.68	36	30.88	20.51	41
Inert objects	38.22	24.62	36	65.39	34.44	41
Active objects	77.42	34.39	36	89.76	33.50	41
Focal fish	117.91	60.12	36	130.32	58.05	41

Note. The number of accounts in each category was compared independently.

How did David Brooks get it so wrong?

Basically it was empirically careless “motivated reasoning”,
which is notoriously typical of David Brooks.

But the authors did find another way of looking at the data that’s closer to what he wanted.

If you collapse everything down to two categories —
the moving stuff (“salient”) and the stationary stuff (“field”) —
and you look only the subject of the first sentence of each description,
you find that the Americans did feature the "salient objects" somewhat more often,
while the Japanese did feature the "field" more often:

Table 3
Subject of the First Sentence in Descriptions of Scenes by American and Japanese Participants

	American			Japanese		
Category	*M*	*SD*	*n*	*M*	*SD*	*n*
Salient objects	5.77	2.43	36	3.78	2.74	41
Field	2.14	2.42	36	4.20	2.73	41

Note. The number of accounts in each category was compared independently.

So which table of results should you pick?

Well, it depends on what you want to prove -- your motivation.

…Duh…

And there are publications showing that it's remarkably easy
to turn "Westerners" into "Asians" (or vice versa)
in experiments like this –

…for example, by changing how crowded the pre-experiment waiting room is…

And it’s not just David Brooks who does this kind of thing,
though he happens to be particularly good at finding striking (non-) facts.

The process of social stereotype formation works the same way for all of us,
and it’s remarkably insensitive to real-world statistics,
because it follows the rules of emotionally-loaded confirmation bias.

It works like this:

You consider the hypothesis that people of type X tend to have property P.
You might start with this idea because "everybody knows it",
or because you happen to observe an instance of X exhibiting P.
You have a strong emotional reaction to type X people or to the attribute P or both.
Now when you observe Xs exhibiting P you think to yourself, "Yes! Just as I thought!"

And now the combination of X and P is emotionally loaded, even if it wasn't to start with.

When you come across Xs not exhibiting P, you don't register it to the same degree, because why would you?

When you come across non-Xs exhibiting P, you don’t connect it to your hypothesis, because why would you?

So your belief that people of type X tend to have property P gets gradually stronger.

(Of course, some people just start and end with "everybody knows it", but you're an empirical sort of person…)

Summary:

I've sketched four examples
where "science" is described in potentially misleading ways,
sometimes because of the unavoidable limitations of ordinary language,
and sometimes because of bad faith or motivated reasoning.

But the vast majority of generic statements have no connection to "science" --
they're just regular people expressing facts about the world around us
(or at least what we think are facts).

So let's take a quick look at one of the constructions we use,
namely generic plural nouns --
and what that construction means,
based on the work of the philosopher Sarah-Jane Leslie.

The abstract for a talk Prof. Leslie gave on “Generics and Generalization”:

Generic sentences express generalizations about kinds, such as "tigers are striped", "ducks lay eggs", and "ticks carry Lyme disease". I present and review emerging evidence from adults and children that suggests that generics articulate cognitively default generalizations — i.e., they express basic, early-developing inductive generalizations concerning kinds. Further evidence suggests that these generalizations don't depend solely on information about prevalence. For example, "ticks carry Lyme disease" is accepted, but "books are paperbacks" is not, despite the fact – well-known and acknowledged by participants – that paperbacks are much more prevalent among books than Lyme-disease-carrying is among ticks. Similarly, both adults and preschoolers understand that, e.g., only female ducks lay eggs, yet they are more likely to accept "ducks lay eggs" than "ducks are female". Rather than depending solely on information about prevalence, these primitive generic generalizations are sensitive to a number of content-based factors, such as whether the property in question is dangerous or otherwise striking (as in "ticks carry Lyme disease"), or is an essential or characteristic property of the kind (as in "ducks lay eggs"). This suggests that our most basic means of forming inductive generalizations about kinds is not guided by prevalence alone, but also reflects our nature as learners.

As Prof. Leslie explained in a series of publications,
the relationship between prevalence and generic statements is weak and complex.

Everyone knows that "Ticks carry Lyme Disease",
although only a minority of ticks do so (14% in one study).

Everyone knows that "Mosquitoes carry West Nile Virus",
though the infection rate in the epicenter of a recent epidemic was 0.355%
(and the rate was essentially zero outside of the epicenter).

And, of course, everyone knows that "Ducks lay eggs" and "Lions have manes",
though in each case the prevalence is at most 50%.

In contrast, people generally reject the statement that "Books are paperbacks",
although the majority of books are in fact paperbacks.

She suggests that the "content-based factors" involved in such judgments include:

1. whether the property is perceived to be characteristic of the kind, or an essential property of the kind;
2. whether the property is striking and/or dangerous; and
3. whether the members of the kind who lack the property have an equally salient alternative property.

And this side-steps the whole “natural kinds” question…

Ducks, ticks, mosquitoes, lions, books – OK.

But “Psychopaths”? “Schizophrenics”? “Immigrants”? “Asians”?

”Men” and “Women”?

From an earlier slide:

The limitations of ordinary language in this area
pose difficult problems for scientists, journalists, teachers, and everyone else.

But the problems are especially hard to avoid
for AI researchers aiming to turn large text collections
into understanding of the world that the texts discuss.

So what is to be done?

Some suggestions:

An easy task: Ask the system things like ”Do ducks lay eggs?”
and then ask follow-up questions, to determine what the system “knows”
and how it explains and justifies its answers.

A harder task: Ask the system “What does X say about Y in this op-ed piece?”
and then “What research did they cite?
And did they describe it accurately?”

An even harder task: Ask the system things like
“Summarize and evaluate the social-science literature
on ethnic differences in individualism vs. collectivism”.

**…but at this point we’re starting to imagine a robot scientist --
which would really be “artificial intelligence”!**

State Of The Art -- A conversation with GPT-3:

Me: Do ducks lay eggs?
GPT-3: Yes, ducks lay eggs.
Me: What do you mean by that?
GPT-3: Ducks lay eggs .
Me: Do all ducks lay eggs?
GPT-3: Yes, all ducks lay eggs.
Me: What about lions?
GPT-3: I'm not sure about lions.
Me: Do lions have manes?
GPT-3: I'm not sure about lions.

So maybe there's some headroom even in the easy task?

Me: Where do eggs come from?
GPT-3: Eggs come from chickens. Chickens lay eggs and then farmers collect them to sell.
Me: What about ducks? Do they lay eggs?
GPT-3: Yes, ducks also lay eggs. Duck eggs are often used in Chinese cuisine.
Me: What about roosters? Do they lay eggs?
GPT-3: No, roosters do not lay eggs.
Me: Are roosters chickens?
GPT-3: No, we consider roosters and chickens to be two different types of poultry.
Me: What do you call female chickens?
GPT-3: We call them hens.
Me: What do you call male chickens?
GPT-3: We call them roosters.
Me: And do female chickens lay eggs?
GPT-3: Yes, hens lay eggs.
Me: How about male ducks?
GPT-3: Yes, male ducks also lay eggs.
Me: How about lions?
GPT-3: No, lions do not lay eggs.
Me: Why not?
GPT-3: Lions are mammals, and mammals do not lay eggs. Instead, they give birth to live young.
Me: Are male lions mammals?
GPT-3: Yes, all lions are mammals.
Me: So do male lions also give birth to live young?
GPT-3: No, only female lions give birth to live young. Male lions do not have this capability.
Me: But male ducks lay eggs.
GPT-3: Yes, as we said before, male ducks also lay eggs.