Graduate School Admissions (GSA).[1] A graduate school became suspicious when it noticed a *negative correlation* between being female and being accepted (across *all* applicants in a given year). Further investigation revealed that — *within each department* — there was a *positive* correlation between being female and being accepted!

- Assume *two* departments: E and $\overline{E}$ (think: *E*asy & *not*-*E*asy). Let $M \overset{\text{def}}{=}$ male, $\overline{M} \overset{\text{def}}{=}$ female (gender binary), $A \overset{\text{def}}{=}$ accepted, and $\overline{A} \overset{\text{def}}{=}$ rejected. Finally, assume these *acceptance rates*.

	M	$\overline{M}$	Overall
E	$60/80$	$40/50$	$100/130$
$\overline{E}$	$10/50$	$20/80$	$30/130$
Overall	$70/130$	$60/130$	$130/260$

[1] This is a simplified version [10] of the example originally discussed in [2]. See vudlab.com/simpsons/ for some fun visualizations of this example.

- I'll offer a "rationalizing explanation" of *why* such examples may *seem* "paradoxical" shortly. First, I'll say some things about the existing literature on Simpson's Paradox (SP).
- The existing literature on SP tends to be concerned (almost exclusively) with doing the following two kinds of things.
  1. Giving *mathematical explanations* (or *characterizations*) of the (algebraic) *structure* of SP cases. The 2×2 case (as above), has a simple algebraic characterization [10].[2]
  2. Giving *causal explanations* of the statistical distributions observed in *actual* SP cases [11]. In GSA, the causal story is simple (and benign). Females tend to apply to the "harder" department, *i.e.*, the department with lower acceptance rate.
- My aim is different — I want a "rationalizing explanation" of *why* SP may (reasonably) *seem paradoxical* in the first place.

[2] The probability calculus (a decidable algebraic theory [8]) can be used to provide a *general* characterization of the 2×2 case. See Extras slides 11–12.

- There are various ways to try to explain (or *explain-away*) "paradoxicality." My approach fits the following mold [5].
  - **Step 1**. Identify the (or *an* explanatorily salient) argument form $\mathcal{A}$ that is *invalidated by* the "paradoxical" cases.
  - **Step 2**. Identify a "similar" or "nearby" (more on this below) form of argument form $\mathcal{A}^\star$, which *is universally valid*.
  - **Step 3**. Argue (or invite the listener to consider) that the (reasonable) *appearance* of paradoxicality stems (at least, in part) from *conflating* $\mathcal{A}$ and $\mathcal{A}^\star$ — that is, from hearing examples of the "paradox" as invalidating argument form $\mathcal{A}^\star$ (which *would be* paradoxical, since $\mathcal{A}^\star$ is *valid*).
- Ideally, the argument forms $\mathcal{A}$ and $\mathcal{A}^\star$ should:
  (a) have sufficiently similar *logical forms*, and
  (b) employ sufficiently similar (probabilistic) *concepts*.
- Moreover, ideally, the "explanation" should (c) be *fully general* (*i.e.*, be applicable to *all instances* of the "paradox").

- Carnap [1] distinguished two senses of "confirmation."

Confirmation as firmness. P confirms$_f$ Q, on the (indicative) supposition that R, just in case Q is (sufficiently) *probable*, conditional upon the conjunction $P \& R$. Formally,

$$\boxed{\begin{array}{c} P \text{ confirms}_f\ Q \text{, on the (indicative) supposition that } R \\ \text{iff} \\ \Pr(Q \mid P \& R) > t.^3 \end{array}}$$

Confirmation as increase in firmness. P confirms$_i$ Q, on the (indicative) supposition that R, just in case P and Q are *positively correlated*, conditional upon R. Formally,

$$\boxed{\begin{array}{c} P \text{ confirms}_i\ Q \text{, on the (indicative) supposition that } R \\ \text{iff} \\ \Pr(Q \mid P \& R) > \Pr(Q \mid \overline{P} \& R). \end{array}}$$

- Since our "explanation" of SP will be formal/universal, the *interpretation* of $\Pr(\cdot \mid \cdot)$ can be allowed to vary, as needed.

[3] Here, $t \geq 1/2$ is some (possibly contextually determined) *threshold*.

- That brings us to **Step 1** of our "explanation." Here is the argument form $\mathcal{A}$ that is invalidated by instances of SP.[4]

$$(\mathcal{A})\quad \begin{array}{l} (1)\ P \text{ confirms}_i\ Q\text{, on the supposition that } R. \\ (2)\ P \text{ confirms}_i\ Q\text{, on the supposition that } \overline{R}. \\ \hline \therefore (3)\ P \text{ confirms}_i\ Q\text{, } \textit{unconditionally}.^5 \end{array}$$

- In our GSA example, $\mathcal{A}$ is instantiated as follows:

$$\begin{array}{l} (1)\ \overline{M} \text{ confirms}_i\ A\text{, on the supposition that } E. \\ (2)\ \overline{M} \text{ confirms}_i\ A\text{, on the supposition that } \overline{E}. \\ \hline \therefore (3)\ \overline{M} \text{ confirms}_i\ A\text{, } \textit{unconditionally}. \end{array}$$

- The Pr-distribution determined by our GSA 2×2 table above constitutes a counterexample to the validity of $\mathcal{A}$.

[4]Strictly speaking, SP is more general than $\mathcal{A}$, since (i) it can also involve *dis*confirmation$_i$ and/or *irrelevance*, and (ii) it can involve random variables with *more than two values*. My explanation(s) go through in full generality.

[5]Note: "unconditionally" just means "*on a tautological supposition*."

- To see this, we can add *probabilistic labels* to our GSA table.

	M	$\overline{M}$	Overall
E	$60/80 = \Pr(A \mid E \,\&\, M)$	$40/50 = \Pr(A \mid E \,\&\, \overline{M})$	$100/130 = \Pr(A \mid E)$
$\overline{E}$	$10/50 = \Pr(A \mid \overline{E} \,\&\, M)$	$20/80 = \Pr(A \mid \overline{E} \,\&\, \overline{M})$	$30/130 = \Pr(A \mid \overline{E})$
Overall	$70/130 = \Pr(A \mid M)$	$60/130 = \Pr(A \mid \overline{M})$	$130/260 = \Pr(A)$

- With these probabilistic labels in place, we can now see that:
  (1) $\Pr(A \mid E \,\&\, M) < \Pr(A \mid E \,\&\, \overline{M})$. That is, $\overline{M}$ confirms$_i$ A, on the supposition that E. In words: being female is *positively correlated* with acceptance, *in department* E.
  (2) $\Pr(A \mid \overline{E} \,\&\, M) < \Pr(A \mid \overline{E} \,\&\, \overline{M})$. That is, $\overline{M}$ confirms$_i$ A, on the supposition that $\overline{E}$. In words: being female is *positively correlated* with acceptance, *in department* $\overline{E}$.
  (3) $\Pr(A \mid M) > \Pr(A \mid \overline{M})$. That is, $\overline{M}$ *dis*confirms$_i$ A, *unconditionally*. In words: being female is *negatively correlated* with acceptance, *in the general population*.
- This explains, purely in confirmation$_i$ terms, what SP *is*.

- So much for **Step 1**. On to **Step 2**. I will examine *two* confirmation-theoretic (and 1 causal) "explanations" of SP.
- **First Way**. This way involves a postulated simple conflation of Carnap's two senses of confirmation (note: this is a well-established human psychological conflation [3]).

$$(\mathcal{A}_1^\star)\quad \begin{array}{l} (1)\ P \text{ confirms}_f\ Q\text{, on the supposition that } R. \\ (2)\ P \text{ confirms}_f\ Q\text{, on the supposition that } \overline{R}. \\ \hline \therefore (3)\ P \text{ confirms}_f\ Q\text{, } \textit{unconditionally}. \end{array}$$

- $\mathcal{A}_1^\star$ is *universally valid*. This is easy to show, since it (essentially) boils down to the following "most" validity.[6]

$$(\mathcal{A}_1^\star)\quad \begin{array}{l} (1)\ \text{Most } P \,\&\, R\text{-worlds are } Q\text{-worlds.} \\ (2)\ \text{Most } P \,\&\, \overline{R}\text{-worlds are } Q\text{-worlds.} \\ \hline \therefore (3)\ \text{Most } P\text{-worlds are } Q\text{-worlds.} \end{array}$$

[6]See Extras slide 13 for a rigorous algebraic proof of the validity of $\mathcal{A}_1^\star$.

- **Second Way**. This way does not involve a simple conflation of Carnap's two senses of "confirms." It involves *only* confirmation as increase in firmness (confirms$_i$).
- This time, the conflation will involve what I will call *suppositional vs conjunctive* confirmation.
- Here's an analogy to help get a grip on the distinction. Consider the following two indicative conditional forms:
  (I) If R, then if P then Q.
  (II) If $P \,\&\, R$, then Q.
- Many philosophers [9] have claimed that (I) and (II) are (in general) *equivalent*.[7] As such, many think conflating (I) and (II) is OK. Here's an analogous *confirmational* pair.
  ($\mathbb{S}$) P confirms Q, *on the supposition that* R.
  ($\mathbb{C}$) $P \,\&\, R$ confirms Q, *unconditionally*.

[7]This equivalence is called *import-export*. See [7, 4, 6] for discussion.

- Interestingly, $\mathbb{S}$ and $\mathbb{C}$ are *equivalent* for confirms$_f$, but *not* for confirms$_i$. To see the former, simply apply the definition of confirms$_f$ above (then the equivalence will be obvious).
- The fact that $\mathbb{S}$ and $\mathbb{C}$ are *not* equivalent for confirms$_i$ is the key to our Second Way. To wit, here's my second $\mathcal{A}^\star$ form.

$$(\mathcal{A}_2^\star)\quad \begin{array}{l} (1)\ P \,\&\, R \text{ confirms } Q, \textit{unconditionally}. \\ (2)\ P \,\&\, \overline{R} \text{ confirms } Q, \textit{unconditionally}. \\ \hline \therefore (3)\ P \text{ confirms } Q, \textit{unconditionally}. \end{array}$$

☞ $\mathcal{A}_2^\star$ is *valid* for *both* confirms$_f$ *and* confirms$_i$.[8]

- So, if (3) is false, then at least one of (1) and (2) must also be false. In our GSA example, we have the following contrast:

($\mathbb{S}$) $\overline{M}$ confirms$_i$ A, *on the supposition that* $\overline{E}$.

$\neg$($\mathbb{C}$) $\overline{M} \,\&\, \overline{E}$ *does* ***not*** confirm$_i$ A, *unconditionally*.

[8] See Extras slides 13 and 14 for algebraic proofs of these claims.

- How do our Two Ways fare wrt our three *desiderata* (a)–(c)?
- $\mathcal{A}_1^\star$ (a) has *the same* logical form as $\mathcal{A}$, (b) involves *two* confirmational concepts (but those concepts are conflated in other contexts [3, 4]), and (c) covers *all* SP cases.
- $\mathcal{A}_2^\star$ (a) has a *different* logical form than $\mathcal{A}$ (but the two forms are conflated in other contexts [9, 4]), (b) involves *only one* confirmational concept, and (c) covers *all* SP cases.
- Finally, let's consider Pearl's *causal* approach to SP [11], which can be fit into our mold (see Extras slide 15 for details).
- Pearl's $\mathcal{A}^\star$ (a) has *almost* the same logical form as $\mathcal{A}$ (*if* you think of Pearl's $\mathcal{A}^\star$ as an *enthymeme*), but (b) it involves *causal* and not merely probabilistic/confirmational concepts, and (c) because it (*sensu strictu*) requires an extra premise (*viz.*, P and R are "causally independent," in Pearl's sense), it does *not* apply to *all* cases of SP. [Note: I do not mean to deny that Pearl's story has explanatory value.]

- The following *stochastic truth table* [8] represents all possible probability distributions over the $\{P, Q, R\}$ language, *via* the 7 real variables $a, b, c, d, e, f, g \in [0, 1]$.

P	Q	R	$\Pr(\cdot)$
T	T	T	a
T	T	⊥	b
T	⊥	T	c
T	⊥	⊥	d
⊥	T	T	e
⊥	T	⊥	f
⊥	⊥	T	g
⊥	⊥	⊥	$1 - (a + b + c + d + e + f + g)$

- *In general*, an SP reversal (like our GSA reversal) occurs when the following three inequalities are satisfied.

$$\Pr(Q \mid P \,\&\, R) < \Pr(Q \mid \overline{P} \,\&\, R)$$
$$\Pr(Q \mid P \,\&\, \overline{R}) < \Pr(Q \mid \overline{P} \,\&\, \overline{R})$$
$$\Pr(Q \mid P) > \Pr(Q \mid \overline{P})$$

- Algebraically [8], these three SP inequalities become:

$$\frac{a}{a+c} < \frac{e}{e+g}$$
$$\frac{b}{b+d} < \frac{f}{1-(a+b+c+d+e+g)}$$
$$\frac{a+b}{a+b+c+d} > \frac{e+f}{1-(a+b+c+d)}$$

- It is easy to use `PrSAT` [8] to find *instances* of this pattern (or *any* SP pattern). But, giving a general characterization is quite complex (although, in principle, it is decidable).
- In any case, this does constitute a *general*, algebraic characterization of (dichotomous) SP reversals (which subsumes the case involving statistical frequencies, expressible in terms of 2×2 contingency tables).
- The next two slides contain (algebraic) proofs of the validity of $\mathcal{A}_1^\star$ and $\mathcal{A}_2^\star$ (using this same algebraic setup).

- *Probabilistically*, $\mathcal{A}_1^\star$ is (by definition) as follows.

$$(\mathcal{A}_1^\star)\quad \begin{array}{l} (1)\ \Pr(Q \mid P \,\&\, R) > t. \\ (2)\ \Pr(Q \mid P \,\&\, \overline{R}) > t. \\ \hline \therefore (3)\ \Pr(Q \mid P) > t. \end{array}$$

- *Algebraically* (using the above setup [8]), $\mathcal{A}_1^\star$ becomes:

$$(\mathcal{A}_1^\star)\quad \begin{array}{l} (1)\ \frac{a}{a+c} > t. \\ (2)\ \frac{b}{b+d} > t. \\ \hline \therefore (3)\ \frac{a+b}{a+b+c+d} > t. \end{array}$$

- Cross-multiplying (1) & (2) yields:

$$(1)\quad a > (a+c)\cdot t$$
$$(2)\quad b > (b+d)\cdot t$$

- Adding the lhs & rhs of these and collecting t yields (3). □

$$(3)\quad a+b > (a+b+c+d)\cdot t$$

- *Probabilistically*, $\mathcal{A}_2^\star$ is *equivalent to*[9] the following.

$$(\mathcal{A}_2^\star)\quad \begin{array}{l} (1)\ \Pr(Q \mid P \,\&\, R) > \Pr(Q). \\ (2)\ \Pr(Q \mid P \,\&\, \overline{R}) > \Pr(Q). \\ \hline \therefore (3)\ \Pr(Q \mid P) > \Pr(Q). \end{array}$$

- *Algebraically* (using the above setup [8]), $\mathcal{A}_1^\star$ becomes:

$$(\mathcal{A}_2^\star)\quad \begin{array}{l} (1)\ \frac{a}{a+c} > a+b+e+f. \\ (2)\ \frac{b}{b+d} > a+b+e+f. \\ \hline \therefore (3)\ \frac{a+b}{a+b+c+d} > a+b+e+f. \end{array}$$

- Cross-multiplying (1) & (2) yields:

$$(1)\quad a > (a+c)\cdot(a+b+e+f)$$
$$(2)\quad b > (b+d)\cdot(a+b+e+f)$$

- Adding the lhs & rhs of these and collecting t yields (3). □

$$(3)\quad a+b > (a+b+c+d)\cdot(a+b+e+f)$$

[9] This formulation allows us to prove $\mathcal{A}_2^\star$ in *the same way* we proved $\mathcal{A}_1^\star$.

- Here is Pearl's argument form: $\mathcal{A}_{\mathscr{P}}^\star$.

$$(\mathcal{A}_{\mathscr{P}}^\star)\quad \begin{array}{l} (0)\ \Pr(R \mid do(P)) = \Pr(R \mid do(\overline{P})) = \Pr(R). \\ (1)\ do(P) \text{ confirms}_i\ Q\text{, on the supposition that } R. \\ (2)\ do(P) \text{ confirms}_i\ Q\text{, on the supposition that } \overline{R}. \\ \hline \therefore (3)\ do(P) \text{ confirms}_i\ Q\text{, } \textit{unconditionally}. \end{array}$$

- If we think of it as an *enthymeme* — with premise (0) left *unstated* — then its form is *almost*[10] the same as $\mathcal{A}$.
- Premise (0) asserts that P is *casually independent* of R.
- In GSA, premise (0) says that *intervening on an applicant's gender would not affect the probability that she applies to department E, as opposed to department $\overline{E}$.*
- That seems right. But, there are cases in which (0) *fails*, but SP occurs (and can still seem, to some extent, "paradoxical").

[10] In (1)–(3), $do(P)$ needs to be contrasted with $do(\overline{P})$, not $\overline{do(P)}$, and so the actual form of Pearl's $\mathcal{A}_{\mathscr{P}}^\star$ is slightly different than what I've written here.

[1] R. Carnap, *Logical Foundations of Probability*, 1962. ⟨http://fitelson.org/carnap/logical_foundations_of_probability.pdf⟩

[2] N. Cartwright, "Causal laws and effective strategies," 1979. ⟨http://fitelson.org/Cartwright_CLAES.pdf⟩

[3] V. Crupi, B. Fitelson and K. Tentori, "Probability, confirmation, and the conjunction fallacy," 2008. ⟨http://fitelson.org/pccf.pdf⟩

[4] I. Douven, *The Epistemology of Indicative Conditionals*, 2016. ⟨http://tiny.cc/mir8cy⟩

[5] B. Fitelson and J. Hawthorne, "The Wason Task(s) and the Paradox of Confirmation," 2010. ⟨http://fitelson.org/wason.pdf⟩.

[6] B. Fitelson, "Comments on Khoo & Mandelkern," 2017. ⟨http://fitelson.org/kmc.pdf⟩

[7] ———, "Two New(ish) Triviality Results for the Indicative Conditional," 2016. ⟨http://fitelson.org/triviality_handout.pdf⟩

[8] ———, "A decision procedure for probability calculus with applications," 2008. ⟨http://fitelson.org/pm.pdf⟩

[9] A. Gillies, "Indicative Conditionals," 2014. ⟨http://fitelson.org/gillies_ic.pdf⟩

[10] G. Malinas and J. Bigelow, "Simpson's Paradox," 2016. ⟨http://plato.stanford.edu/archives/sum2016/entries/paradox-simpson/⟩.

[11] J. Pearl, "Simpson's Paradox: An Anatomy," 2011. ⟨http://fitelson.org/pearl_sp.pdf⟩