What happened to fallacies?

Wilfrid Hodges
Queen Mary, University of London
October 2005
www.maths.qmul.ac.uk/∼wilfrid/jadavpura.pdf

Western deductive logic up to mid nineteenth century
was based on three main assumptions:

Assumption 1. We can usually recognise
when an argument is good
and when it's bad.
But sometimes we make mistakes about this.
Experts can teach us how to make fewer mistakes.

Assumption 2. Good arguments are good for a reason which can be stated as a rule and applied in other cases.

Assumption 3. Bad arguments are bad for a reason which can be stated as a rule and applied in other cases.
(These rules are often called *fallacies*.)

Preliminary remarks

Assumption 1 implies that valid reasoning is a transferable skill.
Today we rightly suspect such claims with no experimental backing.
My next lecture will discuss a relevant experiment.

The main successes of modern logic are in analysis of mathematical arguments,
and in areas related more to definition than to deduction.

Assumption 2 seems to be true and deep.

Modern logic retains it.

Dignāga recognises it by requiring an argument
to contain an example (*dṛṣṭānta*) showing that
the implication generalises.

Assumption 3 is clearly wrong.
Bad arguments are bad by not being good,
not by obeying some rule of badness.

Dharmakīrti's notion of *hetvābhāsa* probably doesn't
correspond to the western 'fallacy'.
A *hetvābhāsa* is an area in which an argument is weak:
e.g. it makes false assumptions, or doesn't generalise,
or just doesn't connect up.

An early discovery:
Some bad arguments look as if they follow a sound rule.

Many examples.
I take one from Walter Burley (14th century).

The following rule is sound.

> If A then B. If B then C. Therefore if A then C.

Thus:

> If I call you a donkey, I call you an animal.
> If I call you an animal, I say the truth.
> Therefore if I call you a donkey, I say the truth.

Move One: Rewrite to introduce **singular noun phrases**:

If I call you a donkey, I call you an animal.
If I call you an animal, **my statement** is true.
Therefore if I call you a donkey, **my statement** is true.

Now we see that different statements are referred to in the second premise and the conclusion.

Another example:

Americans celebrate the Fourth of July.
My father-in-law's birthday is the Fourth of July.
Therefore Americans celebrate my father-in-law's birthday.

This illustrates the formal rule

$$P(a), a = b \vdash P(b).$$

Move Two: Rewrite to introduce singular noun phrases **naming classes of things**.

The fourth of July is in the class of **days of the year which are American festivals**.
The fourth of July is my father-in-law's birthday.
Therefore my father-in-law's birthday is in the class of **events which Americans celebrate**.

Moves One and Two, pressed home, say:

Translate premises and conclusion of your argument into statements about classes, using no verbs except 'equals', 'is a member of' and 'is a subset of'.

If the translated premises entail the translated conclusion in the calculus of classes,
then your argument is valid.

Some history

Both Ibn Sīnā and Leibniz seem to have considered this kind of paraphrase but rejected it.

In 1827 George Bentham published it as a way of 'more easily detecting [fallacies]'.
He probably took it from unpublished notes of his uncle Jeremy Bentham, c. 1811.

It was taken over and publicised by William Hamilton of Edinburgh (1830s), and by George Boole in 1847.

It was resisted by Gottlob Frege in 1906, on grounds perhaps close to Leibniz'.

In 1936 Alfred Tarski gave it strong support, saying it guarantees the 'infallibility' (*niezawodnowość*) of inference rules.

By Tarski's truth definition (1933, better 1957), an entailment expressed in a formal language of logic translates into a set-theoretic statement.

For propositional logic the set-theoretic statements are in a small decidable fragment of set theory.

For first-order logic the fragment involved is not decidable (by Alonzo Church 1936).

For example a first-order sentence ϕ:

$$\exists x \forall y (P(y) \to R(x, y))$$

translates into a set-theoretic formula $\phi^\star$:

$$\exists x (x \in d \wedge \forall y (y \in p \to \langle x, y \rangle \in r))$$

where the variables p and r correspond to the relation symbols P and R, and d stands for 'domain'.

A *model* of ϕ is a triple $\langle d, p, r\rangle$, where d is a nonempty set,
p is a subset of d and r is a binary relation on d,
and the values d, p, r make $\phi^\star$ true.
A *model* of a set of sentences is a model of all of them.

If ϕ and ψ are first-order sentences, we write

$$\phi \models \psi$$

for the set-theoretic statement that every model of ϕ is a model of ψ.

For example

$$\exists x \forall y(P(y) \to R(x, y)) \models \forall y(P(y) \to \exists x R(x, y))$$

translates into the set-theoretic statement

$$\forall d \forall p \forall r\ (d \neq \emptyset \wedge p \subseteq d \wedge r \subseteq d^2 \to$$
$$(\exists x(x \in d \wedge \forall y(y \in p \to \langle x, y\rangle \in r)) \to$$
$$\forall y(y \in p \to \exists x(x \in d \wedge \langle x, y\rangle \in r))))$$

The statement

$$\phi \models \psi$$

is called a *sequent*. (From Latin, 'thing that follows'.)
It is *valid* if its set-theoretic translation is true.
If not, it is *invalid*.
Here ϕ, ψ can be sentences of any formal language with a truth definition.
Also we can generalise to have
any number of sentences on the left side.

Every argument that can be formalised as a valid sequent is valid.

Remark on notation

The symbol '$\vdash$' (read 'turnstile') is the usual modern notation for 'Therefore' in a formal argument.

The notation '$\models$' is used when we want to emphasise the set-theoretic translation.

For formal arguments in first-order logic,
$\vdash$ and $\models$ express the same relation.
But it's still useful to have a symbol to express the set-theoretic approach.

Unresolved questions:

1. Which statements of set theory are 'infallibly' true?

2. To translate even Boole's examples,
we need to form classes from arbitrary predicates.
The Russell-Zermelo paradox shows that this is not admissible.

Nevertheless the apparatus seems to work well
as a machine for checking arguments.

At least in first-order logic and other logics with completeness theorems,
all **valid** sequents are caught by a general rule:

> A sequent is valid whenever it is provable in a suitable proof calculus.

(Compare Assumption 2.)

For recognising **invalid** sequents we have a few uniform methods, though they are far from covering all cases.
The rest of this lecture describes one such method.
It covers many medieval examples,
but its usefulness in general is still open.

For simplicity we work in first-order logic with the logical symbols

$$\wedge, \neg, \forall, =$$

and no function symbols.

We say that a relation symbol R *occurs positively (negatively)* in a formula ϕ if R has an occurrence in ϕ that is within the scope of an even (odd) number of negation symbols.

For example in

$$\forall x \neg \forall y \neg (R(x,y) \wedge \neg R(x,x) \wedge Q(x))$$

R occurs both positively and negatively, Q only positively.

Let L be a first-order language.

By a *theory pair* in L we mean an ordered pair $\langle \Phi, \Psi \rangle$ where Φ and Ψ are sets of sentences of L.

We say that this pair is *inconsistent* if there are $\phi_1, \ldots, \phi_m$ in Φ and $\psi_1, \ldots, \psi_n$ in Ψ such that

$$\phi_1 \wedge \ldots \wedge \phi_m \models \neg(\psi_1 \wedge \ldots \wedge \psi_n)$$

is a valid sequent.
(Using the Compactness Theorem this is equivalent to:
$\Phi \cup \Psi$ has no model.)

We say that $\langle \Phi, \Psi \rangle$ is *Arnauld-inconsistent* if
$\langle \Phi', \Psi \rangle$ is inconsistent,
where Φ' comes from Φ by replacing
each relation symbol R in sentences of Φ
by a new relation symbol R',
uniformly throughout Φ.

We say that a pair is *(Arnauld)-consistent* if it's
not (Arnauld)-inconsistent.

Antoine Arnauld,
France 1612–1694.
Wrote the *Port-Royal Logic*
in 1662 with Pierre Nicole

If $=$ is not used,
then in an Arnauld-inconsistent pair $\langle \Phi, \Psi \rangle$ at least one
of Φ, Ψ must already be inconsistent on its own.

Even with $=$,
we can often see by inspection that a pair is
Arnauld-consistent.

Theorem Suppose each relation symbol occurs only
positively or only negatively in sentences of $\Phi \cup \Psi$.
Then if $\langle \Phi, \Psi \rangle$ is inconsistent,
it must be Arnauld-inconsistent.

Proof We assume that $\langle \Phi, \Psi \rangle$ is Arnauld-consistent,
and we expand it, keeping it Arnauld-consistent,
until it's clear that $\Phi \cup \Psi$ has a model.

More precisely we define $\Phi = \Phi_0 \subseteq \Phi_1 \subseteq \ldots$ and
$\Psi = \Psi_0 \subseteq \Psi_1 \subseteq \ldots$, defining Φ_i, Ψ_i by induction on i.

Case One: $\phi_1 \wedge \phi_2$ occurs in Φ.
Then we add ϕ_1 and ϕ_2 to Φ.
This doesn't add any positive occurrences of a symbol R
if R didn't already occur positively.
Similarly with Ψ.

Case Two: $\neg(\phi_1 \wedge \phi_2)$ occurs in Φ.
We claim we can add at least one of $\neg\phi_1$ and $\neg\phi_2$ to Φ
without creating an Arnauld-inconsistency.

By assumption there is a model of $\Phi' \cup \Psi$,
which is a model of $\neg(\phi' \wedge \psi')$.
So it is a model of $\neg\phi'$ or of $\neg\psi'$.
So we can add either $\neg\phi$ or $\neg\psi$ to Φ without creating an
Arnauld-inconsistency.

Similarly with Ψ.

Case Three: $\forall x\phi(x)$ is in Φ.
Then we claim we can add to Φ each sentence $\phi(c)$,
where c is any constant,
without creating Arnauld-inconsistency.

Again the argument is that $\Phi' \cup \Psi$ has a model, and in this
model every named element satisfies $\phi'(x)$.

Similarly with Ψ.

Case Four: $\neg\forall x\phi(x)$ is in Φ.
In this case we expand the language by adding a new
constant c, and we put $\neg\phi(c)$ in Φ.
The justification is again by considering a model of $\Phi' \cup \Psi$.

Similarly with Ψ.

Case Five: If $\neg\neg\phi$ is in Φ, we can add ϕ to Φ, and similarly
with Ψ.

Case Six: If $c = d$ is in Φ or Ψ, and $\phi(c)$ is in Φ, then we can add $\phi(d)$ to Φ.

Similarly with Ψ.

Case Seven: For each constant c in the language, $c = c$ is in both Φ and Ψ.

Claim: There is no atomic sentence $R(c_1, \ldots, c_n)$ such that it's in Φ but its negation is in Ψ, or vice versa.

Proof of claim: No relation symbol occurs negatively in Φ and positively in Ψ, or vice versa.
Our additions have not affected this assumption.

We may need infinitely many additions.
But if each $\langle \Phi_i, \Psi_i \rangle$ is Arnauld-consistent, then so is

$$\langle \bigcup_{i<\omega} \Phi_i, \bigcup_{i<\omega} \Psi_i \rangle.$$

So we can take unions and continue.

Let $\langle \Phi^+, \Psi^+ \rangle$ be the result of making all these additions to $\langle \Phi, \Psi \rangle$.

Put $T = \Phi^+ \cup \Psi^+$.

Then T has the following properties:

- If $\chi_1 \wedge \chi_2$ is in T then χ_1, χ_2 are both in T.
- If $\neg(\chi_1 \wedge \chi_2)$ is in T then at least one of $\neg\chi_1$, $\neg\chi_2$ is in T.
- Similarly with all the other cases.
- If ϕ is an atomic sentence in T then $\neg\phi$ is not in T.

Lemma (Jaakko Hintikka). A theory with these properties always has a model.
This lemma completes the proof of the theorem. □

Example

The following can't possibly be valid:

$$\forall x \exists y (R(x,y) \wedge \forall z (Q(z) \rightarrow R(y,z))) \models \forall x \forall y (R(x,y) \rightarrow Q(y)).$$

This is clear by translating away $\exists$ and $\rightarrow$:

$$\forall x \neg \forall y \neg (R(x,y) \wedge \forall z \neg (Q(z) \wedge \neg R(y,z))) \models$$
$$\neg \neg \forall x \forall y \neg (R(x,y) \wedge \neg Q(y)).$$

Writing this as $\phi \models \neg\psi$, R has only positive occurrences in ϕ and ψ, while Q has only negative.

So by the theorem, the following sequent must also be valid:

$$\forall x \exists y (S(x,y) \wedge \forall z (P(z) \rightarrow S(y,z))) \models \forall x \forall y (R(x,y) \rightarrow Q(y)).$$

Since $=$ doesn't occur, it follows that either (1)

$$\forall x \exists y (S(x,y) \wedge \forall z (P(z) \rightarrow S(y,z)))$$

has no models, or (2)

$$\forall x \forall y (R(x,y) \rightarrow Q(y))$$

is true under all interpretations.

(2) doesn't hold: for example in the natural numbers take $R(x,y)$ to mean $x = y$ and $Q(x)$ to mean $x = 0$. Neither does (1) (Exercise).

A subtler version of the theorem, with a similar proof, says:

If particular relation symbols occur only positively / negatively,
then we can alter just these symbols in Φ and not Ψ.

This in turn is a special case of:

Lyndon Interpolation Theorem Suppose ϕ and ψ are first-order sentences and $\phi \vdash \psi$.
Then there is a first-order sentence θ such that

- $\phi \vdash \theta$ and $\theta \vdash \psi$,
- every relation symbol with a positive occurrence in θ has positive occurrences in both ϕ and ψ,
- every relation symbol with a negative occurrence in θ has negative occurrences in both ϕ and ψ.

Tarski on infallibility:
Alfred Tarski, ‘On the concept of logical consequence’, in Alfred Tarski, *Logic, Semantics, Metamathematics: papers from 1923 to 1938*, ed. John Corcoran, Hackett Publishing Company, Indianapolis, Indiana 1983 pp. 409–420.

Tarski’s model-theoretic truth definition:
Alfred Tarski and Robert Vaught, ‘Arithmetical extensions of relational systems’, *Compositio Mathematica* 13 (1957) 81–102.

Lyndon’s interpolation theorem:
Roger C. Lyndon, ‘An interpolation theorem in the predicate calculus’, *Pacific Journal of Mathematics* 9 (1959) 129–142.
Lyndon’s proof is proof-theoretic. The model-theoretic proof above uses a lemma of Hintikka reported in §2.3 of:
Wilfrid Hodges, *A Shorter Model Theory*, Cambridge University Press, Cambridge 1997.

Medieval material:
Wilfrid Hodges, ‘Detecting the logical content: Burley’s “Purity of Logic” ’, in *We Will Show Them! Essays in Honour of Dov Gabbay*, ed. Sergei Artemov et al., College Publications, London 2005, Volume 2 pp. 69–115.

Víctor Manuel Sánchez Valencia, *Studies on Natural Logic and Categorial Grammar*, PhD dissertation, University of Amsterdam 1991.