A plausible intelligence is a fallible intelligence

Seminario de Inteligencia Artificial y Lógica

26 de agosto de 2025

Raymundo Morado

Computing Machinery and Intelligence, 1950

- Q: Add 34957 to 70764.
- A: (Pause about 30 seconds and then give as answer) 105621.

Doug meets Nicolai

- Is there anything you specialize in?
- WHAT DOES "SPECIALIZE" MEAN? PLEASE USE A COMPLETE SENTENCE.
- When were you written?
- LAST COMPILATION: "4/27/83 19.247"
- What are feet?
- FEET ARE 12.0 INCHES.
- What are arms?
- THAT INFORMATION IS CLASSIFIED.

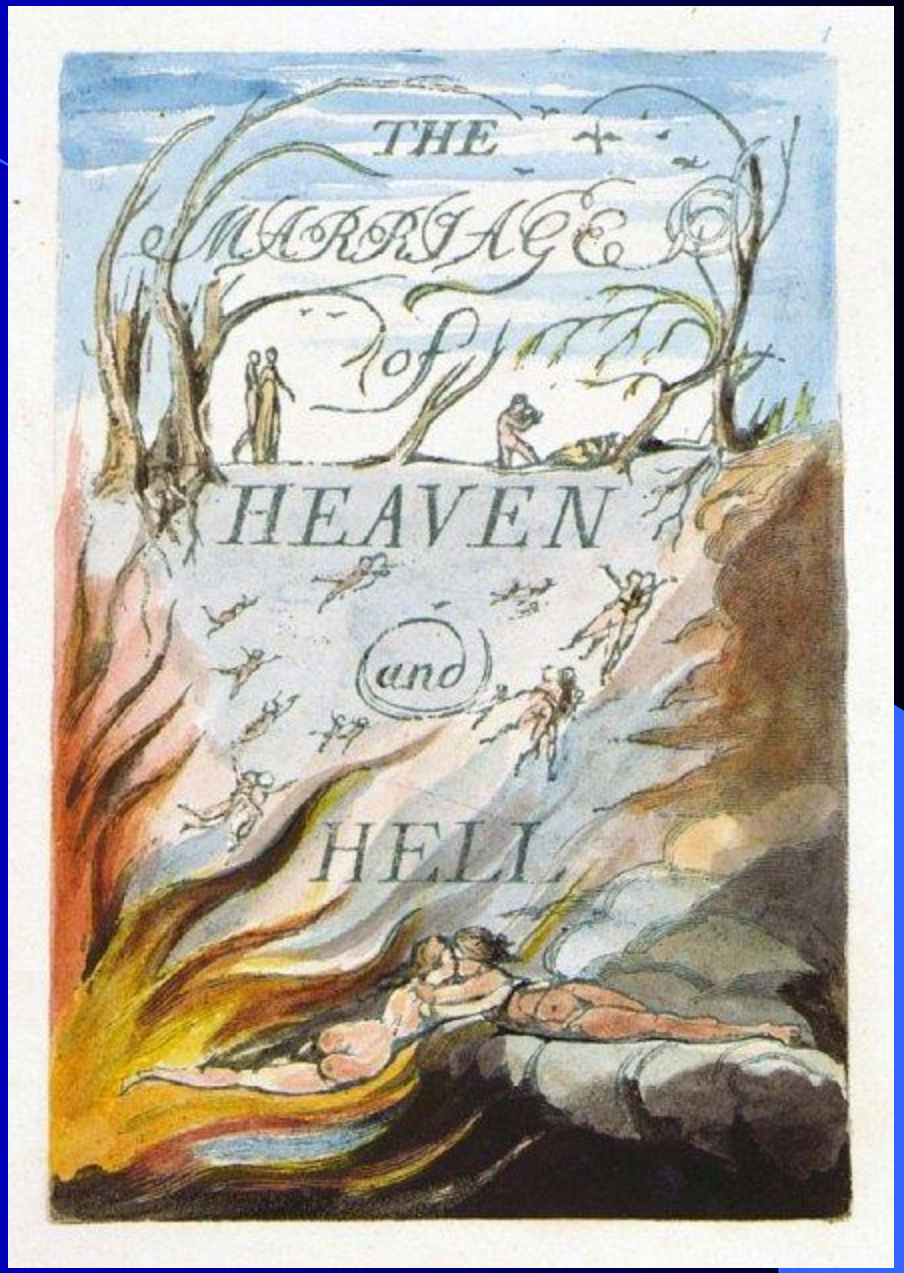
- What is your opinion of M. C. Escher?
- M. C. ESCHER, ARTIST, TWENTIETH CENTURY
- Have you seen the latest issue of "Scientific American"?
- I DO NOT HAVE ANY "SCIENTIFIC AMERICAN".

- ARE YOU "(32) DOUGLAS R. HOFSTADTER"?
- One and the same.
- (83) "PARSING ERROR IN PREVIOUS INPUT"
- How many letters are there in an ordinary envelope?
- 18.
- HOW MANY COMPUTER PROGAMMERS DOES IT TAKE TO SCREW IN A LIGHT BULB?
- 4.
- INCORRECT. THAT IS A HARDWARE PROBLEM.
- What is a computer progammer?

- WOULD YOU LIKE ME TO EXPLAIN THE JOKE?
- What does "to screw" mean?
- IN WHAT CONTEXT?
- In bed.
- "TO SCREW IN BED" MEANS "TO HAVE SEX".
- What is having sex all about?
- I CAN RECOMMEND AN EXCELLENT BOOK ON THAT SUBJECT.
- Tell me, tell me!
- THE BOOK IS CALLED "GODEL, ESCHER, BACH".

"Eternity is in love with the productions of time"

- William Blake, *The Marriage of Heaven and Hell*, 1790.



Ambrose Bierce (1906):

Logic, n. The art of thinking and reasoning in strict accordance with the limitations and incapacities of the human misunderstanding.

Classical logic is an unattainable idealization

- Methods in CS to recover from inconsistencies including. chronological backtracking, dependency-directed backtracking and truth maintenance.
- *Reductio ad absurdum* depends on our notion of an absurdity.

Absurdity, n. A statement or belief manifestly inconsistent with one's own opinion. Bierce (1906).

- We learn to live with errors, redistribute workloads and establish emergency mechanisms to ensure a graceful degradation of output if we cannot mask the errors.
- The fact that we are fallible does not mean that we cannot be rational if by *rationality* we mean the ability to change our beliefs in the presence of evidence in a sensible way.
- `The horse raced past the barn fell".

Reductio ad Absurdum and *Modus Tollendo Tollens*

- Classical logic offers RAA and MTT.
- The problem is that as it is normally proposed, what we obtain from a conflict is the addition of a negation, not the retraction of an affirmation.
- "defeasible reasoning supports alternative and mutually exclusive conclusions drawn from incomplete information". Besnard.

Fallible inference as moral fault

- Arnauld and Nicole (1662): “il y a une infinité d' esprits grossiers et stupides que l' on ne peut reformer en leur donnant l' intelligence de la verité, mais en les retenant dans les choses qui sont à leur portée, et les empeschant de iuger de ce qu'ils ne sont pas capables de connoître”.

- ``Sometimes we contradict ourselves; this is true in our ordinary life, and it is mere prejudice which supposes that we may never do so meaningfully in our logical and mathematical life". Meyer, Routley, Dunn.
- ``im Unbewuβten die Gedanken besonders bequem nebeneinander wohnen, auch Gegensätze sich ohne Widerstreit vertragen, was ja oft genug auch noch in Bewuβten so bleibt". Freud

- The idea of a universe that contains real inconsistencies has always fascinated us. From Cratilus or Heraclitus to Priest or Peña.
- It has even been said that modern science shows how the principle of noncontradiction holds for relatively stable processes, but not for fully dynamic ones, since these are continuously becoming something else, ``revealing their alterity'' (de Gortari).
- Peirce: "Logic teaches us to expect some residue of dreaminess in the world, and even selfcontradictions".

- Inconsistencies can result because of error, ambiguity, or contradictory sources of information. More interestingly, inconsistency may indirectly result because of an incomplete knowledge of the world.

- In an accounting database one would expect to find the constraint ``the sum of the debits and credits for each account is 0''. Schneider

- Such uses of ``consistency'' are legitimate, and it is useful to generalize the notion of inconsistency to any violation, logical or not, of constraints. This yields at least two important kinds of inconsistencies: semantic and syntactic.

SEMANTIC AND SYNTACTIC CONTRADICTIONS

- Scholars of Names and Debaters:
- Hui Shih (c. 350-305 B. C.) and ``The frog has a tail" and ``The tortoise is longer than the snake".
- Kung-sun Lung (b. 380? B. C.) talks of birds flying into a pool as something self-falsifying or illogical.
- Alice laughed. `There's not use trying,' she said: `one CAN'T believe impossible things.' `I daresay you haven't had much practice,' said the Queen. `When I was your age, I always did it for half-an-hour a day.‘ Carroll.

Absurdity and mere oddity

- The remark of the Queen that Alice found impossible to believe was simply that the Queen was one hundred and one, five months and a day.
  - How quaint the ways of Paradox!
  - At common sense she gaily mocks!
  - Though counting in the usual way,
  - Years twenty-one I've been alive,
  - Yet, reck'ning by my natal day,
  - Yet, reck'ning by my natal day,
  - I am a little boy of five!
- William S. Gilbert and Arthur S. Sullivan

Veridical, or truth-telling, paradoxes

- In the benign sense that material implication paradoxes appear counterintuitive, Verdée and De Bal (2006, p.1) “use the term without negative connotation.”

- There is a temptation to restrict the terms ``logical inconsistency'' or ``contradiction'' to phenomena that can be expressed in such a way that it satisfies syntactic criteria. Often we can reduce the semantic to the syntactic by treating them as ``enthymematic inconsistencies'', in the sense of groups of beliefs that would be syntactically inconsistent if we only made explicit beliefs which are considered obvious or fundamental.

- On the other hand, there is also a temptation to say with Wittgenstein that the only impossibility that exists, whether syntactic or not, is *logical* impossibility. He goes on to say that
- [6.3751] *For example, the simultaneous presence of two colours at the same place in the visual field is impossible, in fact logically impossible, since it is ruled out by the logical structure of colour.*

- Salmerón (1991) offers another example of non-syntactic logical inconsistencies. Following ideas of Hare, Salmerón maintains that the claim of universality of statements in educational theory is a logical thesis, *i.e.*, ``a claim about the meaning of the terms and statements".
- A serious judgment about how education should be in a given situation commits the agent to the same belief whenever the agent is confronted with the same behavior or situation. "Y suponer lo contrario, cuando la situación es similar en los aspectos pertinentes, sería una grave inconsistencia lógica".

PREVENTION

- “Beware of bugs in the above code; I have only proved it correct, not tried it.” Donald Knuth.
- Program verification has its practical problems. According to Hewitt, ``There is no way to prove that the process by which the DEC System-20 evolves will result in new releases with consistent formal descriptions".

DAMAGE CONTROL

I do not believe that consistency is necessary or even desirable in a developing intelligent system. No one is ever completely consistent. What is important is how one handles paradox or conflict, how one learns from mistakes, how one turns aside from suspected inconsistencies.

- Minsky (1974, p. 126).

- Since systems seem to become more distributed as they increase in complexity, the result is that inconsistencies are not likely to fade away as we develop more powerful systems; just the opposite.

- Foucault recalls Saint Jerome's criteria to determine authorship. One of them defines the author as ``a field of conceptual or theoretical coherence".
- *Infallibility seems so contrary to common sense that, as you might suspect, it has been advocated by philosophers*. Johnson-Laird and Byrne.
- Attempts at squaring the circle show that conceivability does not suffice to establish possibility. Preface paradox, Kant's antinomies, the lottery paradox, Frege's naive comprehension axiom, etc.

- Kowalski: “There is only one language suitable for representing information --- whether declarative or procedural --- and that is first-order logic. There is only one intelligent way to process information --- and that is by applying deductive inference methods. The AI community might have realized this sooner if it weren't so insular. The data base community, for example, learned its lesson several years earlier.”

ein Satz um so mehr besagt, je kleiner sein Spielraum ist. Carnap

- Classical logic cannot distinguish between the logical closures of two theories if they contain even the smallest contradiction.
- Classical inference rules trivialize any theory with logical inconsistencies.
- On the other hand (semantically), our understanding of what the theories say vanishes if we follow the ideas of Wittgenstein, Carnap or Popper.

PARACONSISTENCY

- ...we must survive contradictory and erroneous data until it is detected and corrected. In the meantime, we cannot avoid giving erroneous answers to questions that depend on erroneous data, but we should answer other questions correctly. O'Donnell.
- Newton C. A. da Costa proposes building a logic for non-trivial inconsistent theories. This approach promises a better modeling of scientific theories.

- Wimsatt: “Formal models of theoretical structures characteristically start with the assumption that the structures contain no inconsistencies. As a normative ideal, this is fine, but as a description of real scientific theories, it is inadequate. Most or all scientific theories with which I am familiar contain paradoxes and inconsistencies either between theoretical assumptions or between assumptions and data in some combination. (Usually these could be resolved if one knew which of several eminently plausible assumptions to give up, but each appears to have strong support, so they ---and the inconsistencies--- remain.)”

Paraconsistent logics promise help in

a) research into the nature of negation and contradiction

- b) explanation of the abstraction schema in set theory
- c) reconstruction of Hegelian dialectics
- d) reconstruction of Meinong's theory of objects and other non-traditional ontologies
- e) study of strongly inconsistent non-trivial theories
- f) study of vagueness and underdetermination

Shapiro, Wand and Martins

- Shapiro has proposed the use of relevance logics to prevent contradictions from ``polluting the data base with every possible conclusion".
- Shapiro, offers a logic called SWM which includes the rule that if an assertion is known to be inconsistent, it will be flagged and prevented from combining with other assertions.
- We can still reason from it, (*e. g.*, for a *reductio*), but the contradiction remains isolated from other propositions.

DAMAGE REPAIR

- 1988 workshop on the future direction on Data Base Management Systems research:
- *There was also unanimity that an active data base system should have a fairly simple rules system that would have extremely high performance. Questions normally addressed by AI researchers under the rubric of expert systems, (e. g. implementing a theorem prover to prove safety of a rule or mutual consistency of a rule set) should be completely avoided.*

- One participant pointed out that a DBMS could simply fire rules at will and remember DBMS states as rules were activated. If the run time system ever returned to the same state again, then inconsistency or unsafety was present. In such a case, the DBMS need only abort the current transaction to back out of the situation. Simple solutions such as this were universally preferred by the participants to attempts at theorem provers.

- This quick-fix attitude and its disregard for the use of logic in integrity constraints was immediately criticized. It was labeled ``parochial'' and the priorities it embodied disavowed by the Editorial Board of the ACM TODS (Association for Computing Machinery, Transaction on Database Systems). (The Editorial Board included two of the workshop participants.)

DEPENDENCY-DIRECTED BACKTRACKING

- Reasoning about uncertain situations should not itself introduce uncertainty --by shoddy record-keeping. Cohen.
- The problem with chronological backtracking.
- Stallman and Sussman ARS (Antecedent Reasoning System).

TRUTH MAINTENANCE SYSTEMS

- There is no basis for assuming that humans are consistent -- not is there any basic obstacle to making machines use inconsistent forms of reasoning. Minsky.

- Doyle-style TMS.
- Problems and limitations.

- A logical treatment of reasoning conflicts is possible, desirable and feasible. Even for logical contradictions, that paradigmatic form of belief conflict. We can prevent them, survive them and eradicate them. Maybe not all of them, maybe not in the ways sketched above, but work on these subjects should remain the task of an attainable logic.

Dirk Gently's Holistic Detective Agency

There is no such word as `impossible' in my dictionary. In fact, everything between `herring' and `marmalade' appears to be missing.

- Douglas Adams.

Some classical assumptions

Le Vieux Monsieur: C'est très beau, la logique.

Le Logicien: A condition de ne pas en abuser.

Eugène Ionesco, *Le Rhinocèros*.

Some traditional metalogical properties of rational inference

- Infallibility
  - No error or revision (MTT, RAA)
  - Consistency
  - No incomplete, erroneous or expired information
- No resource limitations
  - Logical omniscience
  - Computational complexity
- Context-free rules
  - No space or time

Infallibility

Infallibility seems so contrary to common sense that, as you might suspect, it has been advocated by philosophers. Johnson-Laird and Byrne (1992).

Other forms of reasoning:

- Plausible
- Paraconsistent
- Retractable
- *Prima facie*
- Uncertain
- Common sense
- Typical

Logical Omniscience

"...deductive closure (believing all the deductive consequences of your initial beliefs) is not even a regulative ideal. [...] Deductive closure is not in any sense a good thing for a human being [...]. We do not always, or even usually, start with a set of premises and then frantically deduce everything we can. More to the point, to do that would be bad and stupid to the point of insanity, for we would soon be mugged by reality in some form, probably while performing our umpteenth instance of Ampersand- or Vel-Introduction." William G. Lycan, 1994.

No space or time

- Acontextuality
  - Situated logic: McCarthy, Guha, Barwise
- Instantaneity
  - 4-valued information status, Temporal and Dynamic logics
- Non-spatiality
  - Memory limitations, linear logics

- 1969 Robot After dropping a red block, you assume it is still red.
- *Updating our beliefs in a changing world*
- 1978 Airline You are told that Airline Canada flies from Vancouver to Toronto, Boston and Los Angeles. Asked whether it flies to Toulouse you say no.
- *Incomplete information*

- 1980 Tweety You are told that Tweety is a bird and you conclude that Tweety flies.
- *Defeasible reasoning encompasses non-additive reasoning, commonsense inference, prima facie entailments, and fallible reasoning in general.*

- 1981 Nixon From the fact that Nixon is a Quaker you infer that he is a pacifist. From the fact that he is a Republican you infer that he is not a pacifist.
- *Competing putative conclusions*
- 1986 Coffee You believe that if you put sugar in your coffee, it will taste nice. You then conclude that if you put sugar and diesel oil in your coffee it will taste nice.
- *Theory of conditionals*

Nute (1990):

- A man fell from a plane.
- Fortunately, he was wearing a parachute.
- Unfortunately, the parachute didn't open.
- Fortunately, he fell from the plane at a low altitude over a large haystack.
- Unfortunately, there was a pitchfork in the haystack.
- Fortunately, he missed the pitchfork.
- Unfortunately, he missed the haystack...

Life is fired at you at point blank: when the rock you step on pivots unexpectedly, you have only milliseconds to react. Proving theorems is out of the question.

Agre and Chapman (1987).

A frame is a data-structure for representing a stereotyped situation, like being in a certain kind of living room, or going to a child's birthday party. Attached to each frame are several kinds of information. Some of this information is about how to use the frame. Some is about what one can expect to happen next. Some is about what to do if these expectations are not confirmed.

Much of the phenomenological power of the theory hinges on the inclusion of expectations and other kinds of presumptions. A frame's terminals are normally already filled with "default" assignments. Thus, a frame may contain a great many details whose supposition is not specifically warranted by the situation. These have many uses in representing general information, most likely cases, techniques for bypassing "logic," and ways to make useful generalizations.

The default assignments are attached loosely to their terminals, so that they can be easily displaced by new items that fit better the current situation. They thus can serve also as "variables" or as special cases for "reasoning by example," or as "textbook cases," and often make the use of logical quantifiers unnecessary.

MONOTONICITY: Even if we formulate relevancy restrictions, logistic systems have a problem in using them. In any logistic system, all the axioms are necessarily "permissive" -- they all help to permit new inferences to be drawn. Each added axiom means more theorems, none can disappear. There simply is no direct way to add information to tell such the system about kinds of conclusions that should not be drawn! To put

- Retractable reasoning is not necessarily irrational, nor requieres wrong conclusions or uncertain premiases.
- It does not arise from the fact that rules can be revised or from having tacit premises.
- It is not a matter of a true, unspoken background, but is a context-dependent inference that can be blocked.

Some kinds of non-monotonic reasoning

- XVII century:
  - *Induction (Bacon)*
  - *Probabilities (Pascal, Fermat)*
  - *Statistics (Graunt, de Moivre)*
- XIX century:
  - *Abduction (Peirce)*
- XX century:
  - *Closed world assumption, Default reasoning (Reiter)*
  - *Circumscription (McCarthy, Lifschitz)*
  - *Autoepistemic reasoning (Moore, Konolige)*
  - *Inheritance hierarchies (Etherington, Touretzky)*

Raymond Reiter (1939-2002), 1980 "A logic for default reasoning".

John McCarthy (1927-2011), 1980 "Circumscription: A form of non-monotonic reasoning".

Robert C. Moore (1948-), 1985 "Semantical considerations on nonmonotonic logic".

1990 Sarit Kraus, Daniel Lehmann, Menachem Magidor, “Nonmonotonic Reasoning, Preferential Models and Cumulative Logics”.

The facets of rationality :

- constructing a value system,
- acting based on beliefs and desires,
- being logical, etc.

What Is It to Be Logical?

Logicality includes knowledge, abilities and attitudes.

- Logical acuity (inferential ability)
- When to construct, how to offer, and how to evaluate reasons
- How to organize a discussion and accept logical consequences
- Readiness to take context into account
- Ability to recognize the logical structure of an argumentation
- Disposition to look for alternatives

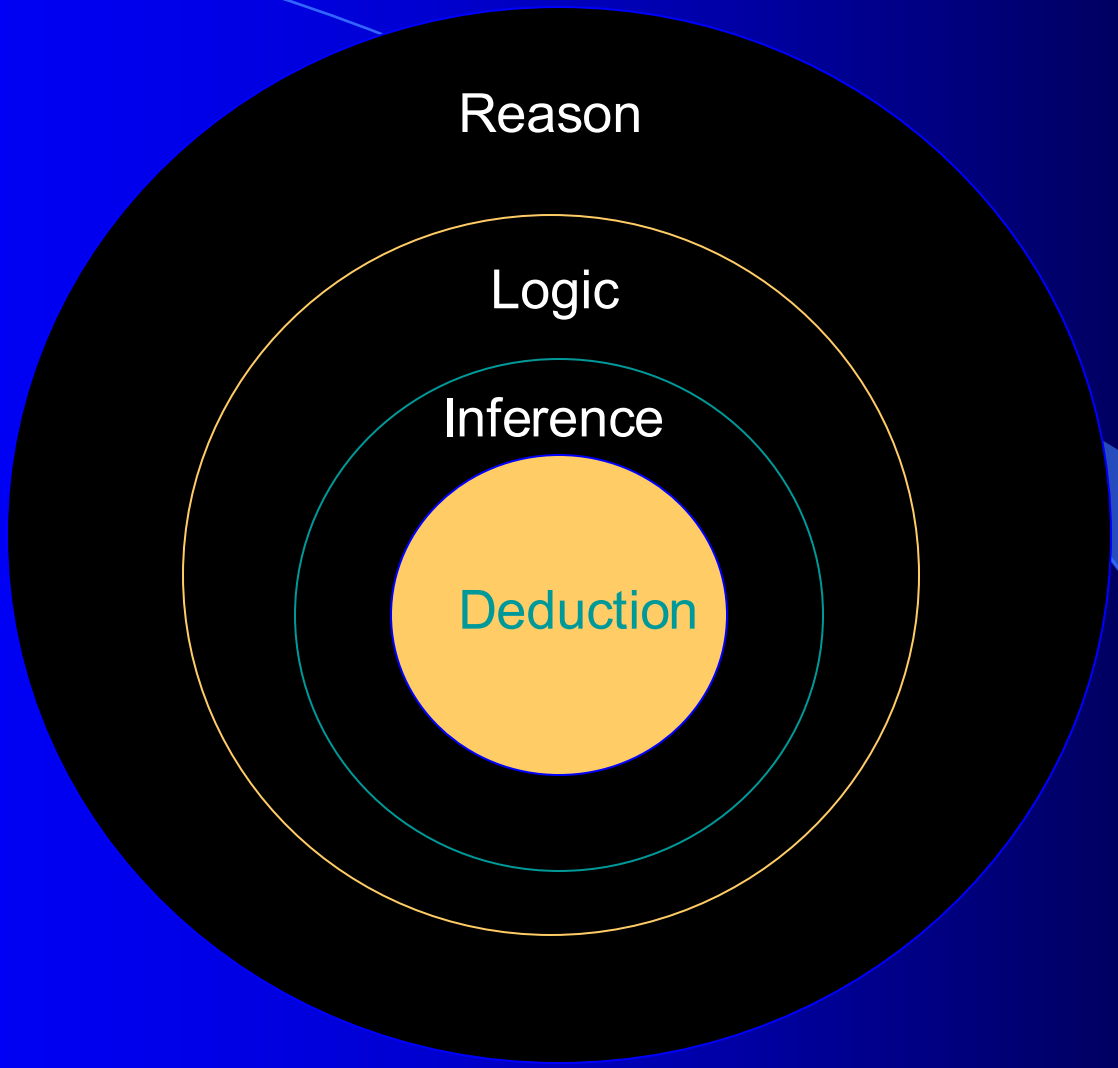
The facets of being logical:

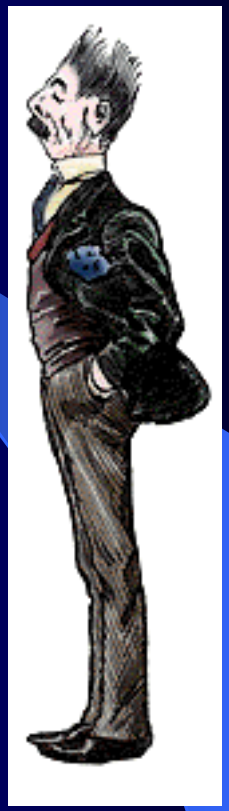
- adopting a belief,
- drawing conclusions from it,
- making plans,
- discerning alternatives,
- discarding the irrelevant,
- arguing properly,
- negotiating,
- understanding arguments from different points of view,
- engaging in counterfactual reasoning,
- evaluating evidence,
- accepting obvious consequences, etc.

We could expand the notion of logicality in at least three ways:

- First, to allow for degrees of logicality and to be able to say that, other things equal, a certain inferential behavior or lack of it is more logical than another. We need to consider both the properties of the context and the existence of cognitive limitations.

- Secondly, incorporating heuristical inferences and in general non-deductive logical structures.
- Thirdly, including non-inferential logical abilities. For instance, the skill to know when to apply the inferential rules, or the ability to inmediately recognize logical truths.
- This notion of logicality is more in accord with the historical use of the term till the mid XIX century.


Reason
Logic
Inference
Deduction



Ought implies can

- To withhold our inference until a complete description of the universe is available would be fatal. In such cases the unreasonable behavior might be not to infer.
- So, we need models that incorporate the provisional status of our inferred beliefs. We can even make the normative claim that for an agent with cognitive limitations to be rational, some of its conclusions must be retractable.
- A model for rationality that does not countenance retractability, a purely monotonic model, fails this norm.

Thank you!