The Philosophy and Ethics of Artificial Intelligence

Overview of Course

- Introduction
- Intelligence
- Consciousness
- Reasoning and Communication
- Ethics and Morality
- Algorithms
- For good or bad ? AI in Medicine, AI in War
- Superintelligence and the Value Loading Problem
- AI and Human Society
- Review/Revision

Reasoning and Communication Paradigms

- The major AI paradigms
- Limitations of paradigms
- Argumentation and Communication
- Argumentation and Epistemology (Extra non-assessed material)

Sources

- Stanford Encyclopedia of Philosophy
- P. M. Dung. On the acceptability of arguments and its fundamental role in non-monotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77(2):321–358, 1995
- S, Modgil. Dialogical Scaffolding for Human and Artificial Agent Reasoning. In: *5th International Workshop on Artificial Intelligence and Cognition*, 73-86, 2017.

Logical (Symbolic) Approaches to AI

Logic and AI

- Early logic-based approaches to AI essentially involved reasoning with symbolic formulae representing beliefs, desires, goals, etc) enabling reasoning about the way the world is (*epistemic reasoning*) and reasoning about what one should do (*practical reasoning*)

- The description of the world is in some logical language (analogous to use of natural language to describe states of affairs, goals, actions) and inference rules are applied to derive new facts about the world, and appropriate actions to achieve goals

- To understand the role of logic in AI let us first briefly review the historical use of logic in mathematics

Logic and the Formalisation of Mathematics

- Mathematical crisis at beginning of 20th century, regarding foundations of mathematics, triggered by discovery of various paradoxes

➔ Research program to prove truths in mathematical fields from given sets of axioms, and showing that each such system is complete (all true statements are provable) and consistent, until Gödel's incompleteness theorem showed this was impossible

- Start with set of self-evidently true axioms (e.g., $\alpha \vee \neg\alpha$) and self-evident logical inference rules (e.g. $\dfrac{\alpha \vee \beta, \neg\alpha}{\beta}$)
- the truth of the axioms and applying inference rules until complex statement in question is deduced.

Logic and the Formalisation of Mathematics

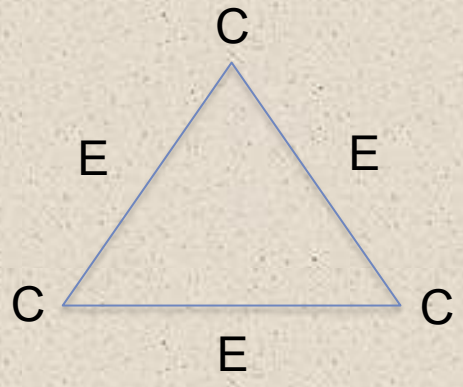
- For example Euclid proved all truths of Euclidean geometry from 5 basic self evident axioms/truths of 2D planes

- Take a simpler example:

  The model/world (mathematical field) we are interested in is a triangle and we can specify a language of symbols E,C representing the set of edges and corners of the triangle

Axiomatic system (self-evident truths)

1. Every member of C is contained in exactly two members of E.
2. Every member of E contains exactly two members of C
3. Every two members of E share exactly one member of C.



One can state these axioms using symbols and logical connectives and quantifiers
Then use rules of logical inference to infer/derive that every three members of E contain exactly three members of C

From Logic in Mathematics to Logic in AI

- The rules of logical inference comprise a proof system (syntax) and the triangle is the model (semantics)
- If anything that can be inferred from the axioms is a true statement about the model, then the proof system is *sound*. If any true statement about the model can be inferred from the axioms (the proof system is *complete*).

- 'Good old fashioned AI' (GOFAI) adopted a similar methodology

Given a model *W* of the world, designer encodes axiomatic truths Δ_0 about *W*

$$W \models \Delta_0$$

in the logical *knowledge base* KB_{Δ_0} of agent (the agent's *beliefs*) and defines a (hopefully) *sound* and *complete* proof system (*PS*) for inferring more complex truths

$$KB_{\Delta_0} \vdash_{PS} \alpha \quad \textit{iff} \quad W \models \alpha$$

For example *PS* = natural deduction and W = interpretation (assignment of truth values to propositional variables)

Logic in AI

Given its goals and what agent believes to be true about the world the agent then **acts**, which results in changes to the world to obtain a new world W` and then **senses** the new facts $\Delta_1$ about the world W` so as to update its knowledge base

Ideally

$$KB_{\Delta_0 \cup \Delta_1} \mid\!\sim_{PS} \alpha \quad \textit{iff} \quad W' \mid\!= \alpha$$

Given what the agent now believes about the world (all the α s is can infer) and its goals the agent then **acts** and senses again, and so the cycle continues

Can logic provide a complete account of all cognition ?

Two *reductionist* arguments:

1) Complete proof system for just first-order (predicate) logic can simulate all of Turing-level computation (i.e., all computable functions, which as we mentioned in lecture on intelligence, implies achievement of any complex goal)

2) If intelligence can be formalised mathematically, and since mathematics can be given logical foundations, then intelligence can be formalised in logic

Monotonic and Non-monotonic Logics

A key difference between the use of logic in mathematics and AI is that in mathematics model/world is fixed/unchanging and axioms describing (limited) world can be specified without worrying about inferences that may later be contradicted when adding more facts (axioms)

In our triangle example, initial 3 axioms are *comprehensive* – no exceptions – and so we do not expect that adding new true facts conflicts with or invalidates inferences we obtained before adding the facts

Hence *classical* logic is ***monotonic***

$$KB_{\Delta_0} \vdash_{PS} \alpha \quad \text{implies} \quad KB_{\Delta_0 \cup \Delta_1} \vdash_{PS} \alpha$$

But real world way too large and complex to encode all possible exceptions in our beliefs. We typically use rules of thumb and then when we discover exceptions, we may withdraw our beliefs

$$KB_{\Delta_0} \vdash_{PS} f \quad \text{but} \quad KB_{\Delta_0 \cup \Delta_1} \not\vdash_{PS} f$$

where Δ_0 contains the rule that all birds typically fly and Tweety is a bird and so we infer that Tweety flies (f), and Δ_1 contains the newly sensed fact that Tweety is a penguin and so we withdraw the conclusion f and indeed now infer $\neg f$

Monotonic and Non-monotonic Logics

We may also withdraw previous inferences because unlike mathematical models/worlds, the real world changes over time

$$KB_{\Delta_0} \mid\!\sim_{PS} f \quad \text{but} \quad KB_{\Delta_0 \cup \Delta_1} \not\mid\!\sim_{PS} f$$

where Δ_1 contains the sensed **new** fact that Tweety's feet are stuck in cement

- The problem with classical logic is that it is *monotonic* – everything that can be inferred from a set of axioms continues to be inferred after adding to these axioms

The only *completely impractical* solution would be to explicitly list all possible exceptions in the rule. Hence instead of any X that is a bird necessarily must fly

$$\forall(X) bird(X) \rightarrow fly(X)$$

we need the rule

$$\forall(X) bird(X) \wedge \neg penguin(X) \wedge \neg cemented\_feet(X) \wedge ..... \rightarrow fly(X)$$

Note that this problem is closely related to *the Frame Problem* which has been studied in AI and philosophy

Non-monotonic Logics

Hence a number of *non-monotonic* logics were developed in AI. These essentially supplemented classical logic with *defeasible* rules of the form

birds typically/usually (rather than necessarily) fly

or *birds fly unless there is evidence to the contrary*

If we subsequently learn that Tweety is a penguin and so doesn't fly, we then only obtain the inference $\neg$ *fly(Tweety)* which invalidates (is preferred to) the inference *fly(Tweety)* given that penguins are a special subclass of birds

Model Theory/Semantics for Logics

- The model theory (semantics) for first order/predicate classical logic typically consists of an *interpretation* :

  - a domain of individuals *D* 'picked out' by/represented by the *constants* in the language (e.g. *tweety*)

  - n-ary tuples of individuals for which each n-ary *predicate* symbol P is true (e.g. interpretation of *unary* predicate *fly* is {(*tweety*)}

  - for each n-ary *function,* mappings from sets of n individuals to single individuals (e.g., unary function symbol *father_of(sanjay) = prem)*

- The model theory (semantics) for non-monotonic logic consists of *preferred* interpretations (models) = *preferential model semantics*

Given two models in which interpretation of unary predicates *bird* and *penguin* is {(*tweety*)}, the one in which interpretation of *fly* is { } is preferred to one in which interpretation of *fly* is {(*tweety*)}

Other Developments in Logic and AI – Modal Logics

- Modal logic extends classical propositional and predicate (first order) logic with operators expressing *modalities*
- A *modal*—a word expressing a modality—qualifies a statement, e.g. . "Sanjay is happy" qualified by saying "Sanjay is *usually* happy"

- *Alethic modalities in Modal Logic* (modalities of truth - possibility and neccessity)

p = it will rain today
$\Box p$ = it's *necessarily* the case that (in all possible worlds) it will rain today
$\Diamond p$ = *possibly* (in at least one possible world) it will rain today

$\Diamond p = \neg \Box \neg p$ - example of an *interaction* axiom

Deontic modalities in Deontic Logic

Oq = q is obligatory
Pq = q is permitted
Fq = q is forbidden

We will revisit Deontic logic when we address the ethics of *AI*

Beliefs, Desires and Intentions (BDI)

- BDI logics describe the mental attitudes of agents acting in the world
- Based on philosopher Michael Bratman's model of *practical* reasoning ("*Intention, Plans, and Practical Reason". Cambridge Uni Press 1987*)

BDI Architectures and Logics have played a prominent role in implementation of *Autonomous Agents*

Beliefs –beliefs about the world (including agent itself and other agents). Note *belief* rather than *knowledge* as what an agent believes may not necessarily be *true*

Desires – motivational state of agent = what agent would like to accomplish (e.g., to become rich, to go to party)

Goals – desires that are being actively pursued by agent (must be consistent unlike desires)

Intentions – *high level* plans put in place to achieve *chosen* goal

- Can make intuitively reasonable inferences from *axioms* of BDI logics, e.g.

$$Intend\ q \longrightarrow Bel \neg q$$

Machine Learning Approaches to AI

Machine Learning – Spring is Here

- Logic based AI did not give particularly impressive results leading to an "AI Winter"

- In recent years machine learning delivered remarkable results, due to development of new algorithms (back-propagation), multi-layer neural networks, availability of massive training data sets, massive increases in computing power and specialised hardware

  e.g.,

  - Google Deep Mind's *Atari Breakout*, *Alpha Go*, *Alpha Go Zero, Alpha Zero*
  - Translation
  - Language Recognition
  - Autonomous Vehicles
  - Computer vision
  - Medical Image Interpretation
  - Text Generation e.g. GPT3 (Shakespeare, jokes, poetry)
  - Deep Learning Robots

- We are in the midst of an AI Spring. Hence all the media coverage about current and future benefits ***and harms*** of AI

Machine Learning – A Very Brief Review of the Fundamentals

- "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E."

- **Supervised Learning**: builds mathematical model from data containing both the inputs and the desired outputs.
- **Semi-supervised learning**: models built from incomplete training data, where a portion of the sample input doesn't have labels.
- **Unsupervised learning**: model built from data which contains only inputs and no desired output labels. Unsupervised learning algorithms are used to find structure in the data, like grouping or clustering of data points.
- **Reinforcement learning**: algorithms given feedback in the form of positive or negative reinforcement in a dynamic environment. Mimics the process of training animals through punishments and rewards,

Machine Learning – A Brief Review of the Fundamentals

- Performing machine learning involves creating a model, which is trained on some training data and then can process additional data to make predictions.
- Most successful models = artificial neural networks
- Deep Learning ((semi)supervised, unsupervised or reinforcement) using multi-layer neural networks has been responsible for many of the WOW successes in AI

  Each layer learns to transform its input data into a slightly more abstract and composite representation. Importantly, a deep learning process can learn which features to optimally place in which level on its own.

Limitations of Logic based and Machine Learning Approaches

Limitations of Logic-based GOFAI

- Simply too many rules needed to be encoded for system to do anything useful. Symbolic solutions didn't scale up from toy examples (eg Blocks World) to real world examples
- Problem of 'brittleness' – small amount of damage leads to complete failure
- Inherently not designed to do low level sensory tasks – the problem of learning/acquiring data in practical way was not addressed in a serious way
- Related *symbol grounding problem* - the symbolic elements of a representation – the constants, functions, and predicates – are typically hand-crafted (i.e., written out by the designer) rather than grounded in data from the real world.

  Philosophically speaking, their semantics (meaning) rely on meanings in the heads of their designers rather than deriving from a direct connection with the world (recall notion of *intentionality*/*aboutness*)

  Practically speaking, hand-crafted representations cannot capture the rich statistics of real- world perceptual data, cannot support ongoing adaptation to an unknown environment, and are an obvious barrier to full autonomy.

Limitations of Logic-based GOFAI

- Machine learning solved symbol grounding problem problem. Instead of manually encoding hundreds of thousands of facts and rules, these were “automatically extracted” from large datasets.

- Also machine Learning less brittle (retain functionality when small amount of damage)

But ….

Limitations of Machine Learning

- Need very large data sets, but humans learn from small amounts of data – capacities "designed" by evolution (e.g., e.g., young children's grasp of basic physics)

- Until recently very little progress on problem of constructing new representations at levels of abstraction higher than the input vocabulary (although deep learning is promising))

  E.g. in computer vision, learning complex concepts such as Classroom and Cafeteria unnecessarily difficult if agent forced to work from pixels as the input representation; instead the agent needs to be able to form intermediate concepts first, such as Desk and Tray, without explicit human supervision.

- Deep learning thus far has no natural way to deal with hierarchical structure – typically sentences are just sequences of words whereas human languages are hierarchically structured (larger structures are recursively constructed out of smaller components) as complex plans and motor control

- Deep learning struggles with open-ended inference – e.g. subtle difference between "John promised Mary to leave" and "John promised to leave Mary" cannot be used to draw inferences about who is leaving whom, or what is likely to happen next.

- Deep learning thus far has not been well integrated with prior knowledge

- Deep learning thus far cannot inherently distinguish causation from correlation

Limitations of Machine Learning

- **Hence**:

  i) problems with generalization/transfer - trained network performs well on one task but poorly on a new task, even if new task is very similar to the one it was originally trained on

  ii) problems with high level cognitive processes such as planning, causal reasoning, analogical reasoning (reasoning that one case or situation is similar to another), all of which are important features of intelligent behavoiur

Limitations of Machine Learning

- Black box problem: Reasoning processes are opaque, even to designers of systems, and so currently cannot be understood by humans (how and why does it give the result it gives)

- Lack of symbolic/language like explanations of reasoning mean that one also cannot integrate reasoning of humans and machines so that they can jointly reason together

As we will see later in the module, both the above have important ethical implications

Problem of Commonsense Reasoning

- Fundamental problem for both ML and logic based AI is *common-sense reasoning*

For example

- fact that we see a bird, and don't *explicitly* check all possible exceptions before concluding that it most likely flies (we don't explicitly check that its assumed ability to fly is consistent with everything else that we believe)

- To some extent these issues are (*theoretically*) addressed by non-monotonic logics.
- However, these logics are:

  1) complex and often not easily understandable to humans

  2) computationally not feasible if we want to ensure rational outcomes (when inferring that Tweety flies, we need to check that this inference is consistent with everything else that is believed – e.g. that it is not a penguin that its feet are not stuck in cement, and so on, which is computationally extremely demanding)

  3) developed for reasoning by *individual* agents and not by multiple agents reasoning jointly/together

Argumentation and Communication

Argumentation and Communication

However, in the last 10-15 years a new way of doing non-monotonic reasoning – *argumentation* - that is:

- more in line with how humans reason and so more understandable to humans

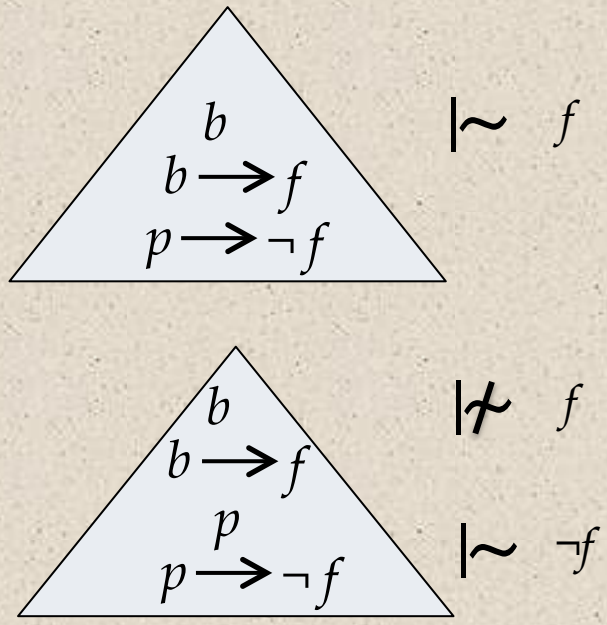
- can be formalised so to give rational outcomes even when computational/cognitive resources are limited

- enables one to integrate the reasoning of multiple agents (human and computer) in the form of dialogues in which agents exchange information and reason together ! (and which will later be shown to be important if want to ensure that AI behaves ethically)

Hence argumentation is a key focus of research in this department

Non-monotonic Logics: a reminder

- Conclusions are withdrawn because new information conflicts with previous conclusions



Since penguins are a special sub-class of birds we prefer inference $\neg f$ over f

Essentially, non-monotonic reasoning is about deciding amongst conflicting inferences that may represent beliefs, goals, or decision options (actions)

Non-monotonic reasoning as argumentation

Rationally choosing amongst

- conflicting beliefs, or

- conflicting desires to adopt as goals (remember that goals, but not desires, must be mutually consistent, i.e,. not conflict with each other), or

- alternative actions for achieving a goal

Is a key aspect of intelligence, and is the central concern of ***argumentation and debate*** - something we as humans are familiar with

Non-monotonic reasoning as argumentation

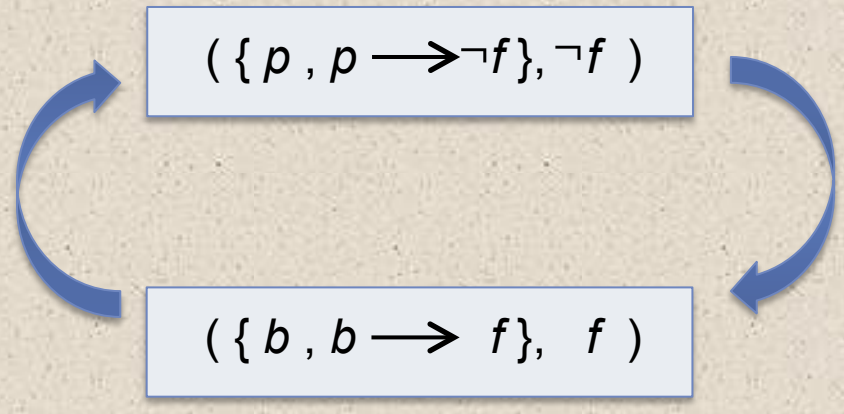
The proof of $\neg f$ from p and $p \rightarrow \neg f$ can be understood as an *argument* consisting of the argument's *grounds/premises* p and $p \rightarrow \neg f$ that support/justify the *claim* $\neg f$

$$(\{ p , p \rightarrow \neg f \} , \neg f)$$

Non-monotonic reasoning as argumentation

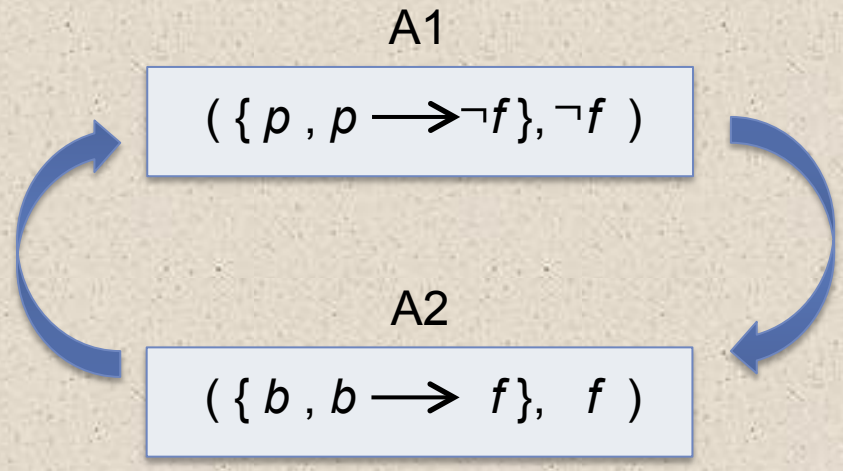
The proof of $\neg f$ from p and $p \rightarrow \neg f$ can be understood as an *argument* consisting of the argument's *grounds/premises* p and $p \rightarrow \neg f$ that support/justify the *claim* $\neg f$

We also have



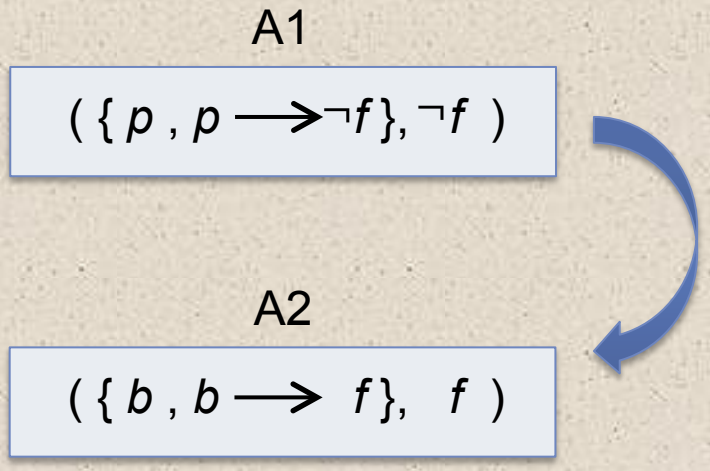
Each argument is a counter-argument to (*attacks*) the other since they have contradictory claims

Non-monotonic reasoning as argumentation



But A1 is preferred to (stronger than) A2 because of the *specificity principle* (properties of subclasses take priority over properties of super-classes) and so A2 cannot *attack* (be moved as a counter-argument to) A1

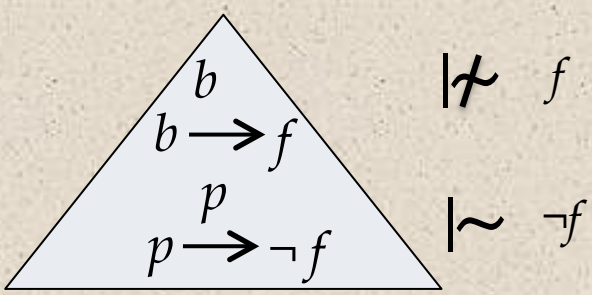
Non-monotonic reasoning as argumentation



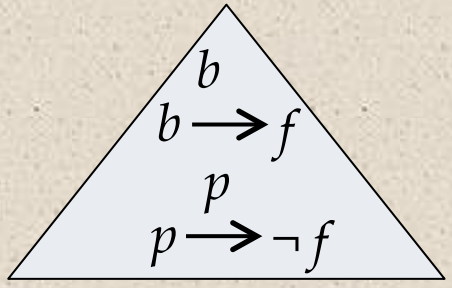
- Hence A1 is the winning argument and the claim ¬ *f* 'wins out' over *f*

Non-monotonic reasoning as argumentation

- In other words we can identify the inferences obtained by proof theory of the non-monotonic logic



By constructing the arguments from



and (possibly using preferences over arguments) determine the winning arguments

Non-monotonic reasoning as argumentation

- We can do this for many non-monotonic logics

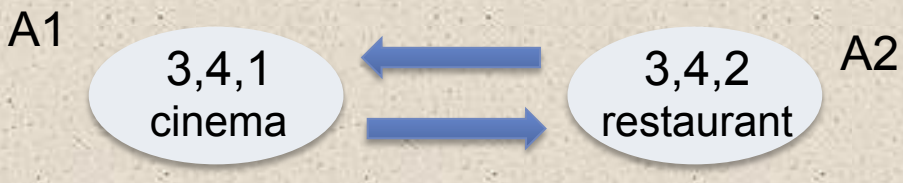
  For example *logic programming* (e.g. Prolog) is a non-monotonic logic

1. Bel *birthday* ∧ ***Des*** *celebrate_birthday* ∧ ***Bel*** *NOT closed_cinema* → ***Goal***(*cinema*)

2. Bel *birthday* ∧ ***Des*** *celebrate_birthday* ∧ ***Bel*** *NOT closed_restaurant* → ***Goal***(*restaurant*)

3. Bel *birthday*

4. Des *celebrate_birthday*



- Neither argument wins out – we are in a genuine dilemma

Non-monotonic reasoning as argumentation

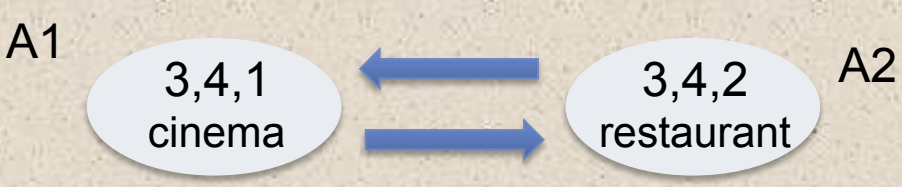
1. Bel *birthday* ∧ ***Des*** *celebrate_birthday* ∧ ***Bel*** *NOT closed_cinema* → ***Goal***(*cinema*)

2. Bel *birthday* ∧ ***Des*** *celebrate_birthday* ∧ ***Bel*** *NOT closed_restaurant* → ***Goal***(*restaurant*)

3. Bel *birthday*

4. Des *celebrate_birthday*

5. Goal(*restaurant*) > ***Goal***(*cinema*)



- Given 5, A2 stronger than (preferred to) A1 and so A1 cannot be moved as a counter-argument to A2. Hence attack from A1 to A2 is cancelled out, and only A2 attacks A1 and so A2 wins

Non-monotonic reasoning as argumentation

1. Bel *birthday* ∧ ***Des*** *celebrate_birthday* ∧ ***Bel*** *NOT closed_cinema* → ***Goal***(*cinema*)

2. Bel *birthday* ∧ ***Des*** *celebrate_birthday* ∧ ***Bel*** *NOT closed_restaurant* → ***Goal***(*restaurant*)

3. Bel *birthday*

4. Des *celebrate_birthday*

5. Goal(*restaurant*) > ***Goal***(*cinema*)

- Given 5, A2 stronger than (preferred to) A1 and so A1 cannot be moved as a counter-argument to A2. Hence attack from A1 to A2 is cancelled out, and only A2 attacks A1 and so A2 wins

Non-monotonic reasoning as argumentation

1. Bel *birthday* ∧ ***Des*** *celebrate_birthday* ∧ ***Bel*** *NOT closed_cinema* → ***Goal***(*cinema*)

2. Bel *birthday* ∧ ***Des*** *celebrate_birthday* ∧ ***Bel*** *NOT closed_restaurant* → ***Goal***(*restaurant*)

3. Bel *birthday*

4. Des *celebrate_birthday*

5. Goal(*restaurant*) > ***Goal***(*cinema*)

6. *closed_restaurant*

- A2 is now attacked by A3 as A2 assumes that *closed_restaurant* is NOT provable. Hence A1 now wins

Non-monotonic reasoning as argumentation

1. p
2. $p \wedge NOT\ q \rightarrow r$

$\mid\sim\ r$

A1

1,2 ✓

Non-monotonic reasoning as argumentation

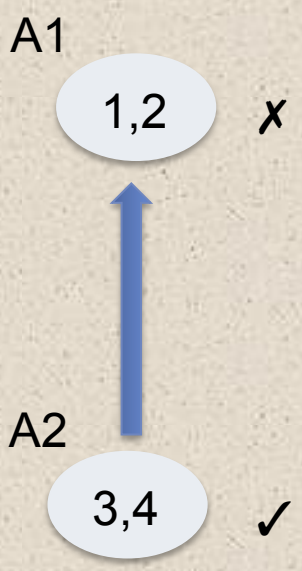
1. p
2. $p \wedge NOT\ q \rightarrow r$

$\mid\sim\ r$

1. p
2. $p \wedge NOT\ q \rightarrow r$
3. s
4. $s \wedge NOT\ g \rightarrow q$

$\not\mid\sim\ r$

$\mid\sim\ q$



Non-monotonic reasoning as argumentation

1. p
2. $p \wedge NOT\ q \rightarrow r$

$\mid\sim\ r$

1. p
2. $p \wedge NOT\ q \rightarrow r$
3. s
4. $s \wedge NOT\ g \rightarrow q$

$\not\mid\sim\ r$

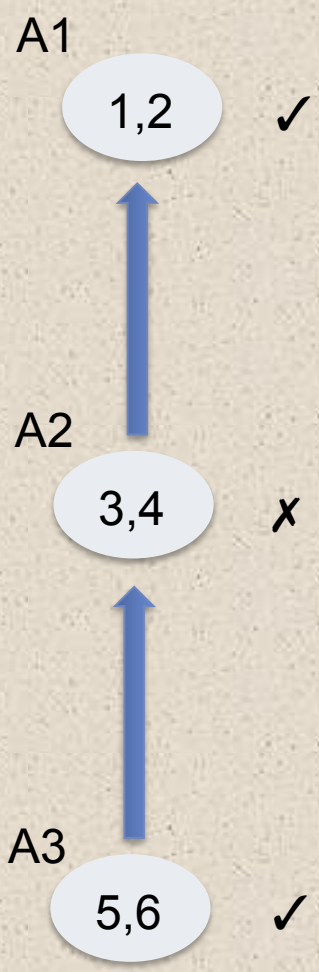
$\mid\sim\ q$

1. p
2. $p \wedge NOT\ q \rightarrow r$
3. s
4. $s \wedge NOT\ g \rightarrow q$
5. f
6. $f \rightarrow g$

$\mid\sim\ r$

$\not\mid\sim\ q$

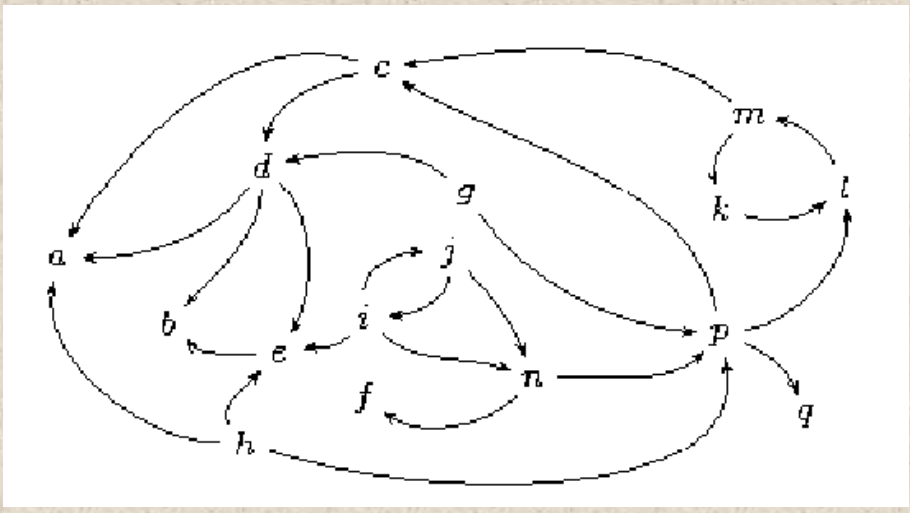
$\mid\sim\ g$



Non-monotonic reasoning as argumentation

"Knowledge Base" containing facts and rules

Argument Framework
$< \mathcal{A}rgs , \mathcal{A}ttacks >$



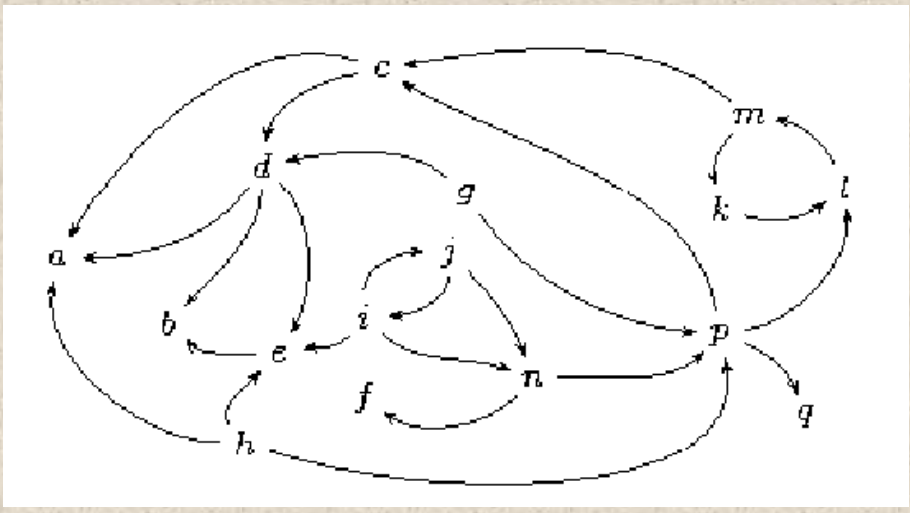
Non-monotonic reasoning as argumentation

"Knowledge Base" containing facts and rules

$|\sim \quad \alpha$

iff

Argument Framework
$< \mathcal{A}rgs\ ,\ \mathcal{A}ttacks >$



α is the claim of a winning argument

Dung's theory of Argumentation [1,2]

Given an Argument Framework < $\mathcal{A}rgs$, $\mathcal{A}ttacks$ > where

$$\mathcal{A}ttacks \subseteq \mathcal{A}rgs \times \mathcal{A}rgs$$

label each argument in $\mathcal{A}rgs$ with either IN (winning), OUT (losing) or UNDEC (undecided)

< $\mathcal{A}rgs$ = {A1,A2,A3}, $\mathcal{A}ttacks$ = { (A3, A2), (A2,A1) } >

1. P. M. Dung. **On the acceptability of arguments and its fundamental role in non-monotonic reasoning, logic programming and n-person games**. *Artificial Intelligence*, 77(2):321–358, 1995

2. M. Caminada. **A Gentle Introduction to Argumentation Semantics**
https://www.semanticscholar.org/paper/A-Gentle-Introduction-to-Argumentation-Semantics-Caminada/21a1a88328f8e274816a70061666ee2ded6f8235

1. p
2. $p \wedge NOT\ q \rightarrow r$

$\mid\sim\ r$

1. p
2. $p \wedge NOT\ q \rightarrow r$
3. s
4. $s \wedge NOT\ g \rightarrow q$

$\mid\not\sim\ r$

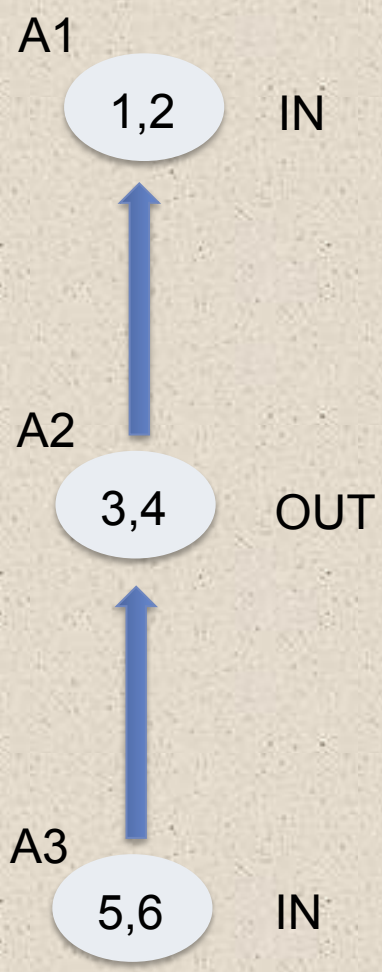
$\mid\sim\ q$

1. p
2. $p \wedge NOT\ q \rightarrow r$
3. s
4. $s \wedge NOT\ g \rightarrow q$
5. f
6. $f \rightarrow g$

$\mid\sim\ r$

$\mid\not\sim\ q$

$\mid\sim\ g$



Dung's theory of Argumentation

Given an Argument Framework < $\mathcal{A}rgs$, $\mathcal{A}ttacks$ > where

$$\mathcal{A}ttacks \subseteq \mathcal{A}rgs \times \mathcal{A}rgs$$

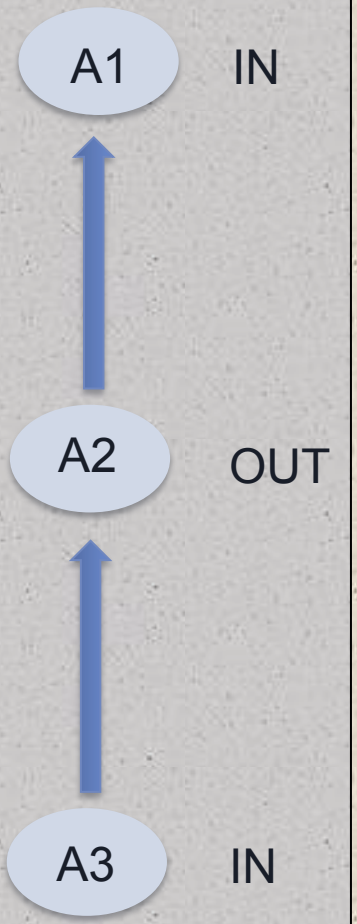
label each argument in $\mathcal{A}rgs$ with either IN (winning), OUT (losing) or UNDEC (undecided)

< $\mathcal{A}rgs$ = {A1,A2,A3}, $\mathcal{A}ttacks$ = { (A3, A2), (A2,A1) } >

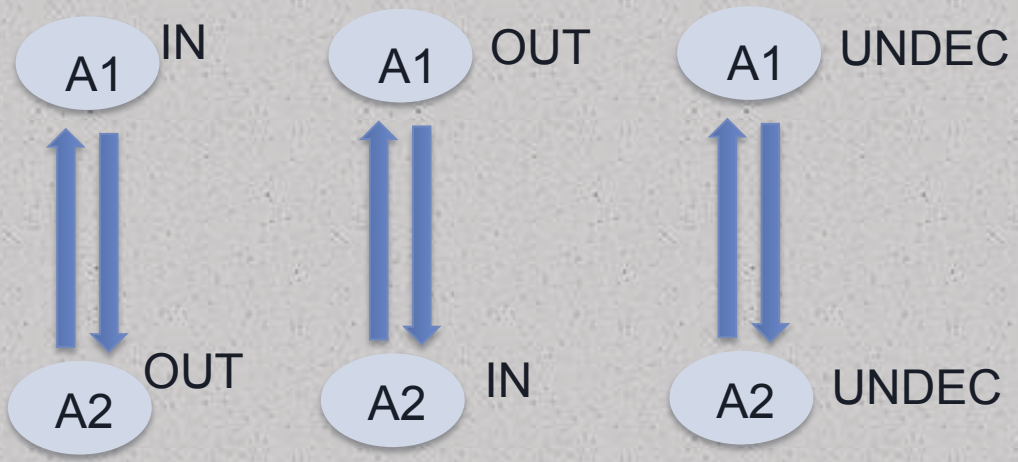
Definition: *A* labelling of $\mathcal{A}rgs$ is a legal (valid) labelling iff

1) If X∈ $\mathcal{A}rgs$ is labelled IN then for every Y such that (Y,X) ∈ $\mathcal{A}ttacks$ Y is labelled OUT
2) If X∈ $\mathcal{A}rgs$ is labelled OUT then there exists a Y such that (Y,X) ∈ $\mathcal{A}ttacks$ and Y is labelled IN
3) If X∈ $\mathcal{A}rgs$ is labelled UNDEC then there exists a Y such that (Y,X) ∈ $\mathcal{A}ttacks$ and Y is labelled UNDEC and there is no Y such that (Y,X) ∈ $\mathcal{A}ttacks$ and Y is labelled is IN

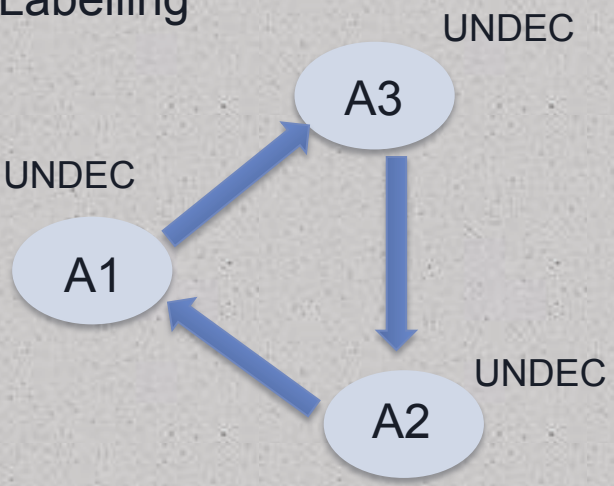
One Legal Labelling …



Three Legal Labellings ….



One Legal Labelling



Dung's theory of Argumentation

Definition: Let $L_1, \ldots L_n$ be the legal labellings of < $\mathcal{Args}$, $\mathcal{Attacks}$ >

1) There will always be a unique L_i that has a minimum number of arguments labelled in IN

$$\exists L_i \forall Lj : IN(L_i) \subseteq IN(Lj)$$

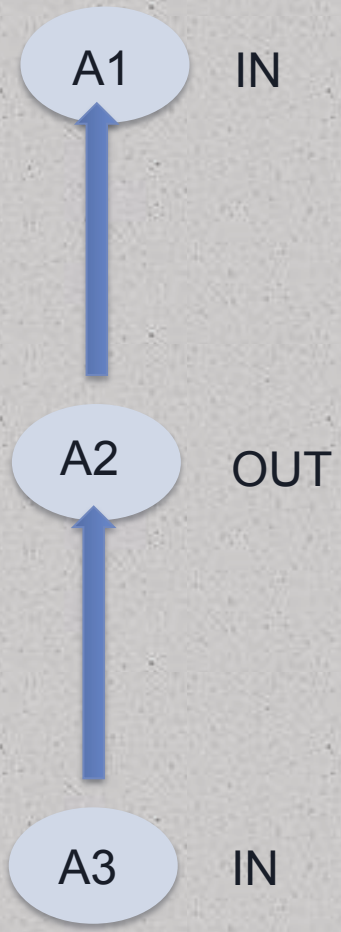
L_i is said to be the ***grounded*** labelling and the arguments labelled IN in L_i (i.e., IN(L_i)) is said to be the grounded extension of < $\mathcal{Args}$, $\mathcal{Attacks}$ >

2) There may be more than one labelling with a maximal number of arguments labelled IN, i.e., the legal labellings L_i such that

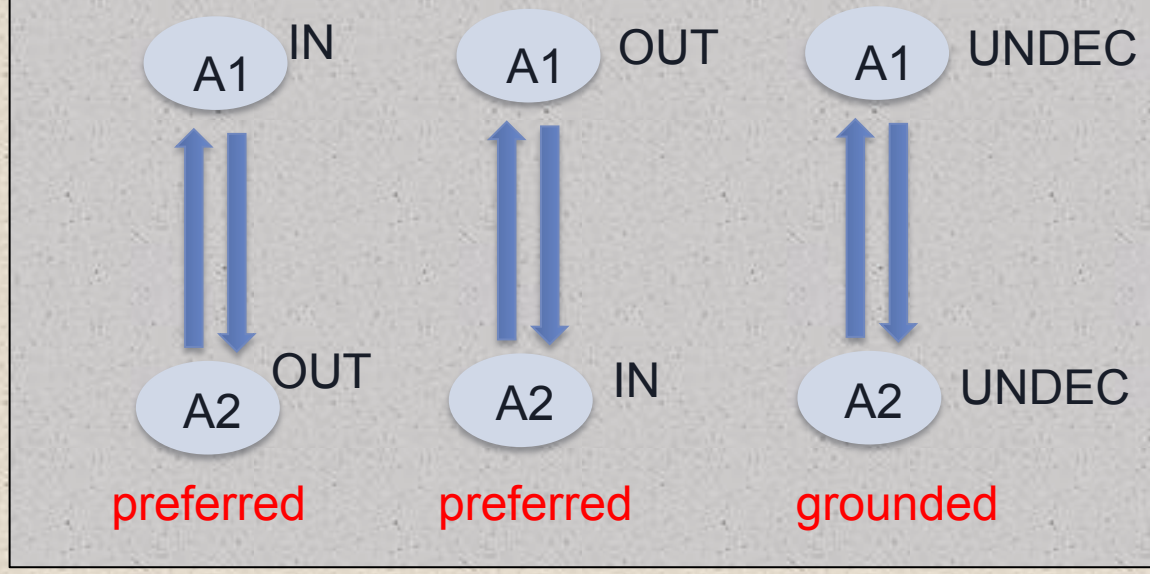
$$\neg\exists Lj : IN(L_i) \subset IN(Lj)$$

Each such L_i is said to be a ***preferred*** labelling and the arguments labelled IN in each such L_i is said to be a preferred extension of < $\mathcal{Args}$, $\mathcal{Attacks}$ >

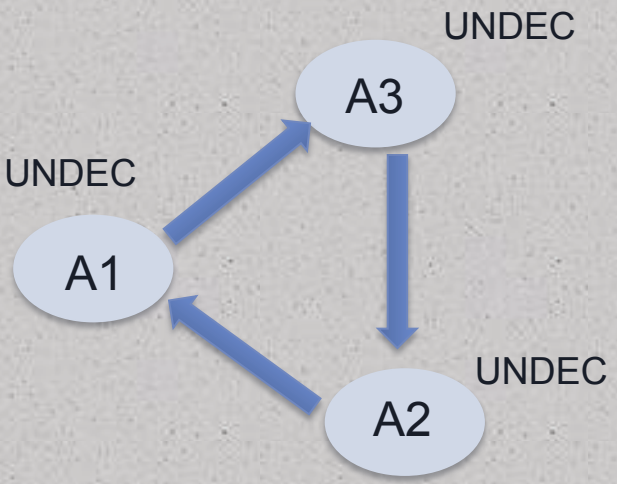
One Legal grounded and preferred labelling



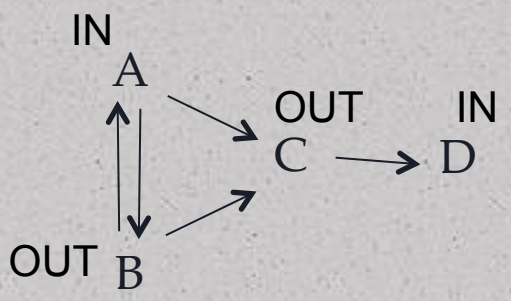
Three Legal Labellings



One Legal grounded and preferred labelling

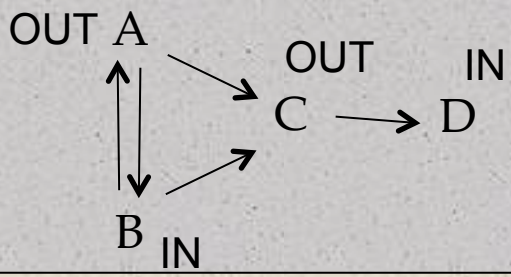


More Examples

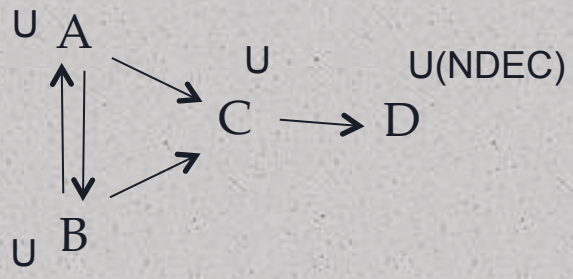


2 *preferred extensions*

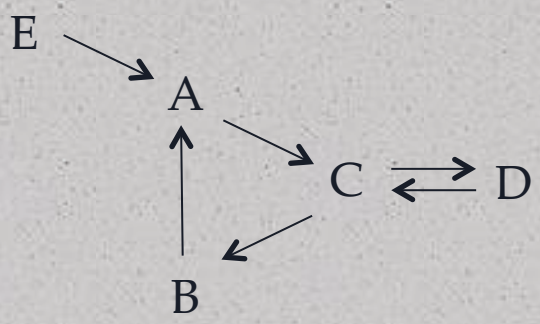
{A,D}



{B,D}



1 *grounded extension* = {}



- {E, C} and {E, D, B} are the two preferred extensions
- {E} is the single grounded extension

(I leave it to you to work out the labellings)

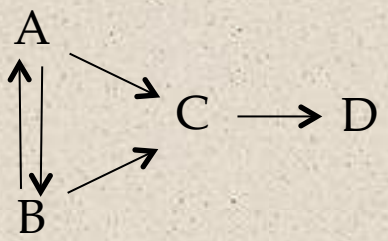
Argument Game Proof Theories

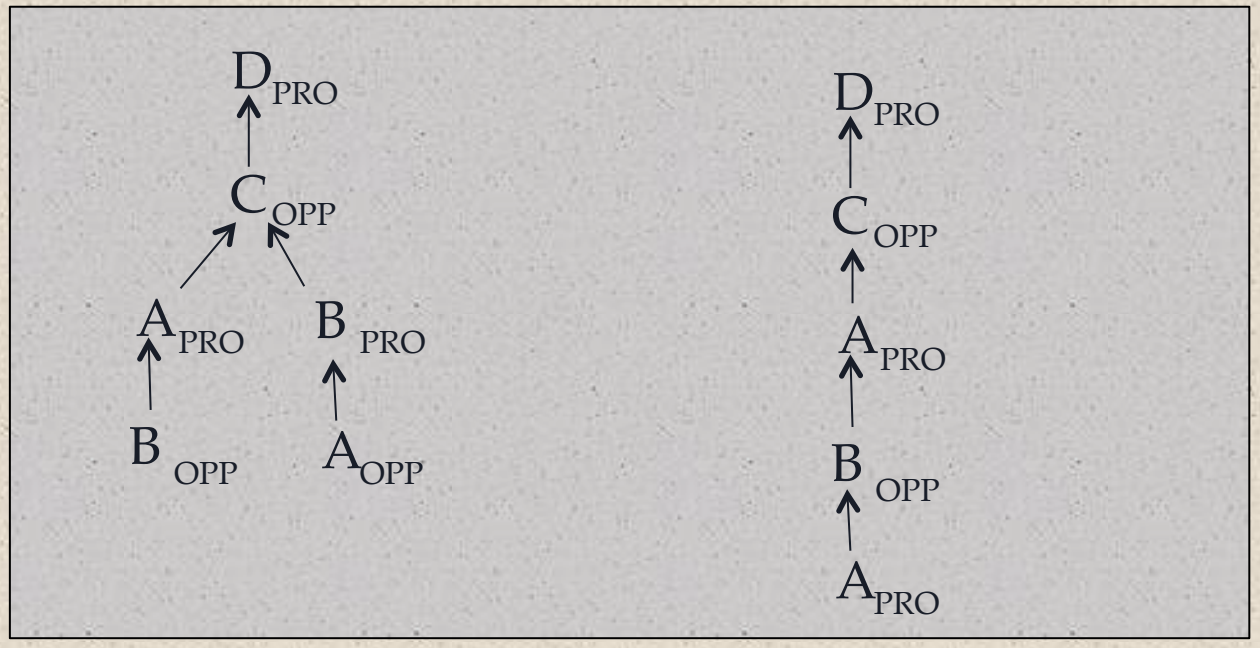
Given argumentation framework < $\mathcal{A}rgs$, $\mathcal{A}ttacks$ > constructed from an agent's knowledge base, argument games defined for deciding whether some $x \in \mathcal{A}rgs$ is in an *s (= grounded or preferred)* extension

Agent plays the roles of PRO and OPP

- PRO moves x
- OPP and PRO then take turns, **referencing** < $\mathcal{A}rgs$, $\mathcal{A}ttacks$ > to move attacking arguments against other player, *subject to the rules of the game that vary according to the criteria/semantics s*
- If PRO successfully counters each OPP move and OPP moves exhaustively (subject to game's rules) then PRO establishes membership of x in an s extension

Argument Game Proof Theories





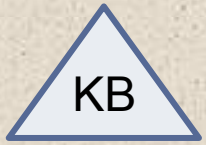
In *grounded* game **PRO cannot repeat i**n any given path from root to leaf (*dispute*) but OPP can repeat in a dispute

PRO loses this game (D is not In the grounded extension)

In *preferred* game **OPP cannot repeat in a dispute but PRO can repeat in a dispute**

PRO wins this game – D is in a preferred extension

Single Agent Non-monotonic reasoning as argumentation



$|\sim \quad \alpha$

iff

Argument Framework
$< \mathcal{A}rgs , \mathcal{A}ttacks >_{\Delta}$

α is the claim of a winning argument under some criteria *s*

- We have assumed a single agent constructing $< \mathcal{A}rgs , \mathcal{A}ttacks >$ from her private KB Δ and using *s* argument game to determine whether an argument is in an *s* extension of $< \mathcal{A}rgs , \mathcal{A}ttacks >$
- But reasoning, especially reasoning about contentious beliefs/goals/actions, is often conducted with others (human and artificial agents) via communicative interactions – through dialogue/debate

Indeed we need to reason ***jointly*** about more complex/difficult issues as we need to integrate individual knowledge and expertise from many sources

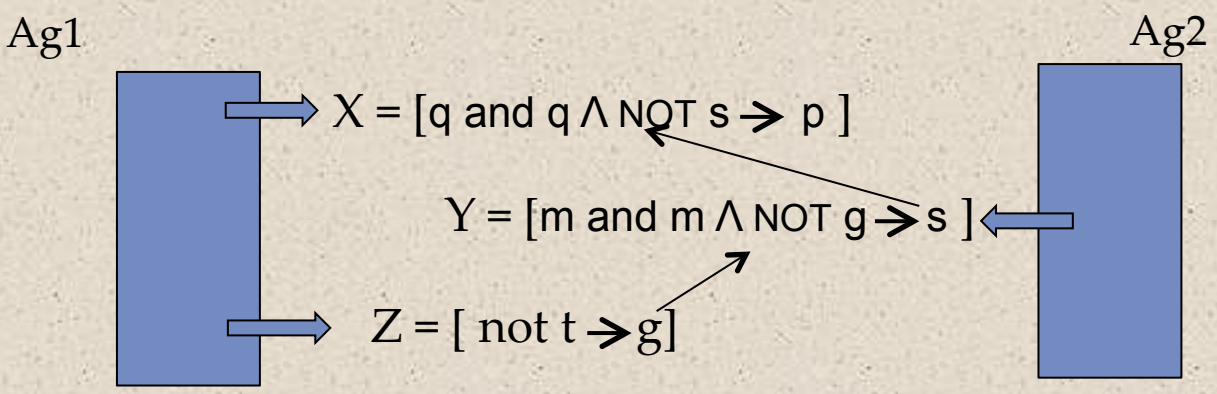
From single agent reasoning to distributed reasoning via dialogue

"The lonesome thinker in an armchair is as marginal as he looks: most of our logical skills are displayed in interaction" – J. Van Bentham

We can adapt argument games to obtain dialogical models of distributed (joint) reasoning.

An *s* dialogue game *with public semantics* is adjudged to be in favor of an argument X (and its claim) iff X is in an *s* extension of the framework built from the contents of what have been publically communicated (and not some private KB)

From single agent reasoning to distributed reasoning via dialogue



Δ incrementally defined by contents of exchanged arguments

q Λ NOT s → p
q
m Λ NOT g → s
m
not t → g

$(\mathcal{A}rgs, \mathcal{A}tt)_{\Delta}$

Ag1 wins grounded dialogue game for argument X claiming that p is true

iff

Argument X claiming p winning under grounded criterion

iff

$\Delta \mid\sim \alpha$

From single agent reasoning to distributed reasoning via dialogue

So a dialogue that reasons that α is the case effectively demonstrates that α can be inferred from the contents of the arguments moved during the dialogue.

An *s (= grounded/preferred)* dialogue game *with public semantics* is adjudged to be in favor of an argument X (and its claim α)

iff

X is in an *s* extension of the argument framework built from the contents Δ of what have been publically communicated in the arguments exchanged during the dialogue (and not the agents' private KBs)

iff

$$\Delta \mid\sim \alpha$$

Argumentation and Communication

We have a way of doing non-monotonic reasoning that is:

- more in line with how humans reason/debate/discuss and so is more understandable to humans
- can be formalised so to give rational outcomes even when computational/ cognitive resources are limited
- enables one to integrate the reasoning of multiple agents (human and computer) in the form of dialogues in which agents exchange information and reason together !
- The reasoning is not taking place within the black box minds of the AIs ! Rather, it is formalised as a publically visible activity, where arguably, reasoning about complex issues (when formal normative models of reasoning are really required) should be transparent (publically visible) and engage multiple minds who bring different knowledge and cognitive capacities to bear
- We will emphasise the importance of these argumentative accounts of reasoning in a later lecture when we discuss joint AI-Human reasoning as being important to ensure that AIs make ethical decisions