Argument, Dialectic and Dialogue: Applications and Theoretical Challenges

Sanjay Modgil

Department of Informatics, King's College London

Outline of Talk

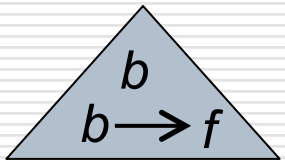
- Non-monotonic Logic, Argumentation and the Dialectical Turn
- Applications
  - Pedagogical
  - The Value Loading (Alignment) Problem
- Theory
  - Towards a Comprehensive Account of Distributed Reasoning
  - Rationality under Resource Bounds

Non-monotonic Logic, Argumentation and the Dialectical Turn

Non-monotonic Logic and Argument

- Non-monotonic logics formalise common-sense reasoning in the presence of incomplete and uncertain information.

- Conclusions are withdrawn because new information conflicts with previous conclusions or the assumptions made in drawing previous conclusions.

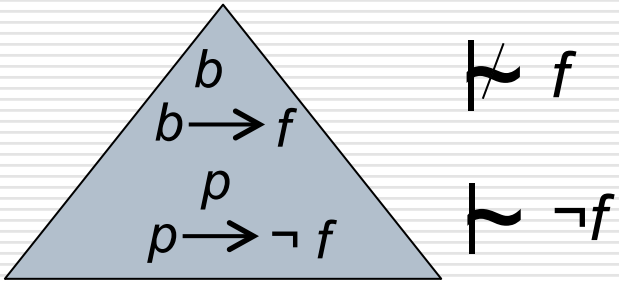


$\mid\!\sim f$

Non-monotonic Logic and Argument

- Non-monotonic logics formalise common-sense reasoning in the presence of incomplete and uncertain information.

- Conclusions are withdrawn because new information conflicts with previous conclusions or the assumptions made in drawing previous conclusions.

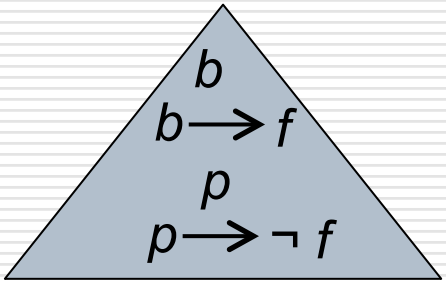


Non-monotonic Logic and Argument

- Non-monotonic logics formalise common-sense reasoning in the presence of incomplete and uncertain information.

- Conclusions are withdrawn because new information conflicts with previous conclusions or the assumptions made in drawing previous conclusions.

- Essentially, *nm* reasoning concerned with arbitrating amongst conflicts; this insight substantiated by *argumentative characterisations of nm inference …*



$\not\vdash\!\!\sim f$

$\vdash\!\!\sim \neg f$

Dung's Argumentation Theory [1]

- Given a set Δ of (ordered) wff in some language $\mathcal{L}$:

1) Construct arguments ($\mathcal{A}rgs$) from Δ

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

Dung's Argumentation Theory [1]

- Given a set Δ of (ordered) wff in some language $\mathcal{L}$:

1) Construct arguments ($\mathcal{A}rgs$) from Δ

(Δ,≤) = ordered set of classical wff and X = $(\Gamma,\alpha)\in\mathcal{A}rgs$ iff Γ is a *consistent minimal* subset of Δ that classically entails α

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

Dung's Argumentation Theory [1]

- Given a set Δ of (ordered) wff in some language $\mathcal{L}$:

1) Construct arguments ($\mathcal{A}rgs$) from Δ

(Δ,≤) = ordered set of classical wff and X = $(\Gamma,\alpha) \in \mathcal{A}rgs$ iff Γ is a *consistent minimal* subset of Δ that classically entails α

2) Define conflict based defeat relation ($\mathcal{D}ef$) amongst $\mathcal{A}rgs$

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

Dung's Argumentation Theory [1]

- Given a set Δ of (ordered) wff in some language $\mathcal{L}$:

1) Construct arguments ($\mathcal{A}rgs$) from Δ

(Δ,≤) = ordered set of classical wff and X = $(\Gamma,\alpha) \in \mathcal{A}rgs$ iff Γ is a *consistent minimal* subset of Δ that classically entails α

2) Define conflict based defeat relation ($\mathcal{D}ef$) amongst $\mathcal{A}rgs$

X = (Γ,α) *defeats* Y = (Π,β) if $\neg\alpha \in \Pi$ and $X \not\prec (\{\neg\alpha\},\neg\alpha)$ where $\prec$ lifts ≤ to sets of formulae

3) Evaluate justified (winning) arguments in directed graph ($\mathcal{A}rgs$, $\mathcal{D}ef$)

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

Logic Programming Example

Δ =

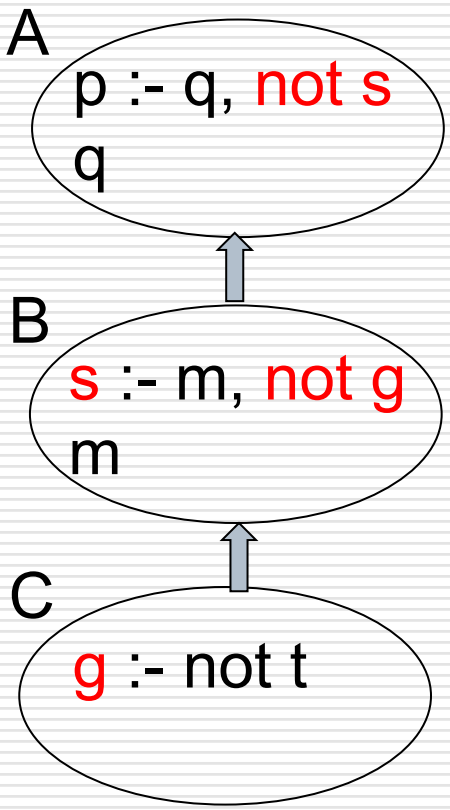
p :- q, not s

q

s :- m, not g

m

g :- not t



Logic Programming Example

Δ =

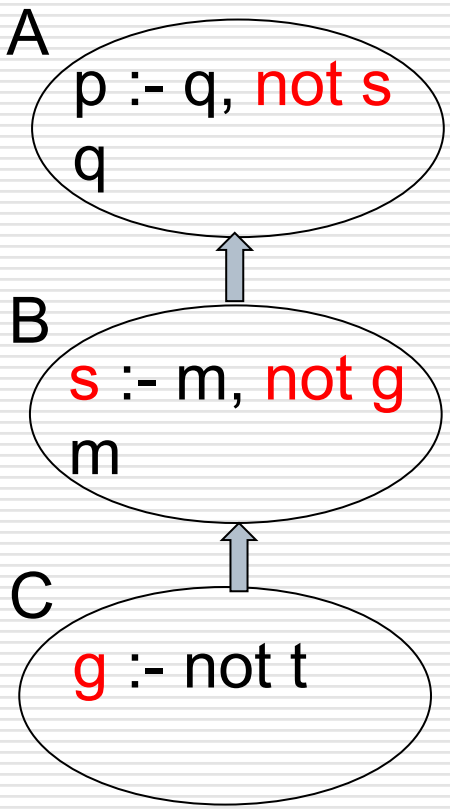
p :- q, not s

q

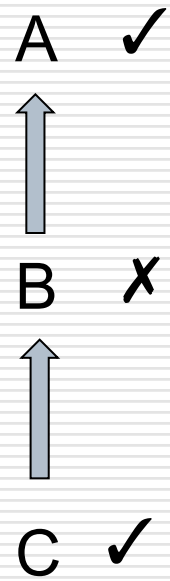
s :- m, not g

m

g :- not t



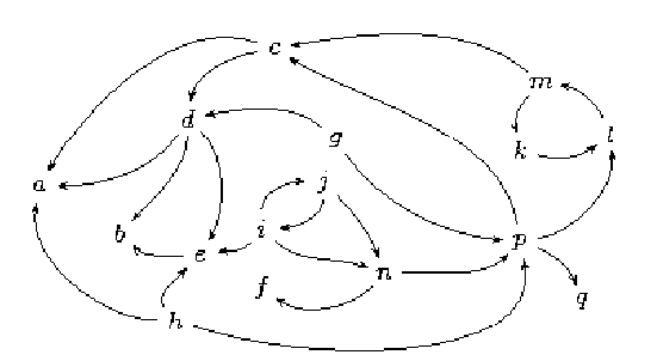
$(\mathcal{A}rgs, \mathcal{D}ef)$



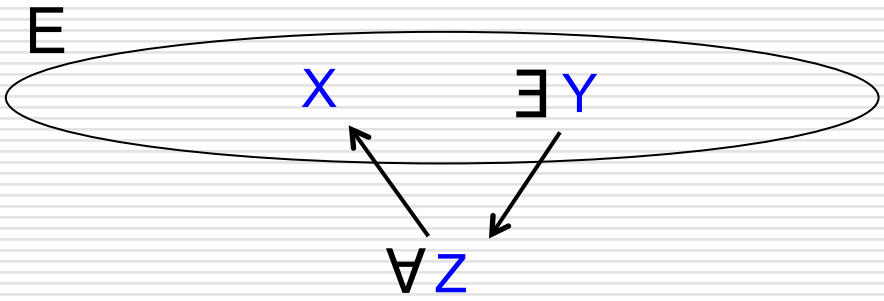
C defends A

Argument Evaluation

- $(\mathcal{Args}, \mathcal{Def})$ defined by (possibly ordered) Δ



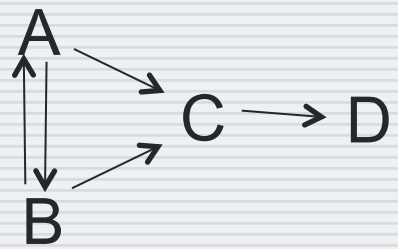
- Sets of *acceptable* (admissible) arguments are **conflict free** and defend themselves against all defeats



Argument Evaluation

- *Maximal* sets of acceptable arguments defined under different semantics

$(\mathcal{A}rgs, \mathcal{D}ef) =$



2 *preferred* ***extensions***

- **{A,D}** (A defends itself against B, and defends D against C)
- **{B,D}** (B defends itself against A, and defends D against C)

Single *grounded* extension is ∅ (arguments cannot self-defend).

(*In general*, the grounded extension is empty if there are no undefeated arguments)

- Many other semantics and developments and variants of Dung's theory

Argumentation-based characterisations of non-monotonic consequence relations

$(\mathcal{A}rgs, \mathcal{D}ef)_{\Delta} \mid\!\sim \alpha$ (the claim of a credulously/sceptically justified argument)

$\Updownarrow$

$\Delta \mid\!\sim_{nml} \alpha$

E.g. *nml* = *Logic Programming*, *Default Logic* (classical logic plus default inference rules), *Prioritised Default Logic*, *Defeasible Logic*, *Preferred Subtheories …*

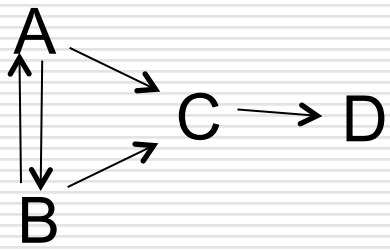
Argument Game Proof Theories

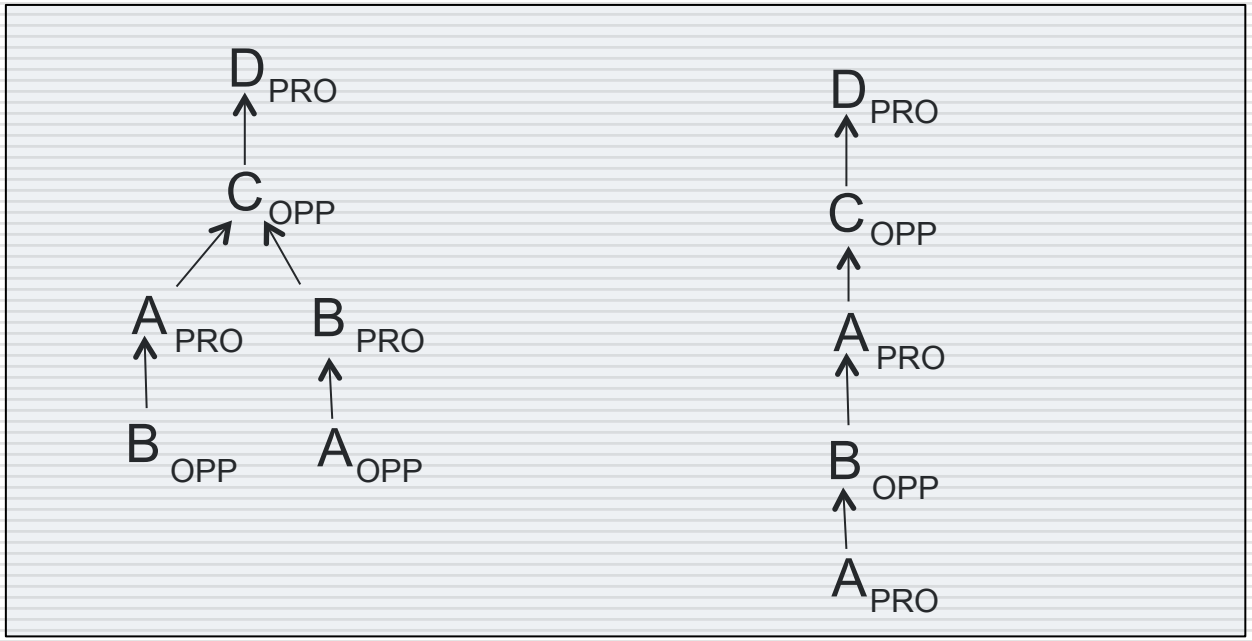
Given argumentation framework (*AF*) $(\mathcal{A}rgs, \mathcal{D}ef)$, argument games defined for deciding whether some $x \in \mathcal{A}rgs$ is in an *s* extension *

- PRO moves x
- OPP and PRO then take turns, **referencing** $(\mathcal{A}rgs, \mathcal{D}ef)$ to move defeating arguments against their counterpart, *subject to the rules of the game that vary according to the 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

* S. Modgil and M. Caminada. *Proof Theories and Algorithms for Abstract Argumentation Frameworks*. In: *Argumentation in AI,* I. Rahwan and G. Simari (eds), 105-132, 2009.

Semantic Specific Argument Game Proof Theories for Deciding Membership of Extensions *





In *grounded* game PRO cannot repeat in a path (PRO loses)

In *preferred* game OPP cannot repeat in a path (PRO wins and is said to have a *winning strategy* = the arguments moved by PRO)

Some Reflections on Proof theory and Model theory

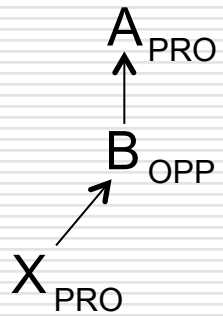
- Argument game proof theories for non-monotonic logics

$$\Delta \mid\!\sim_{nml} \alpha \iff (\mathcal{A}rgs, \mathcal{D}ef)_{\Delta} \mid\!\sim \alpha \iff \exists x \in \mathcal{A}rgs, \text{claim}(x) = \alpha$$

and there is a winning strategy for *x*

- Suppose we don't assume a *given Δ*: agent submits arguments whose contents *incrementally* define Δ (contents of beliefs incrementally reveal themselves)

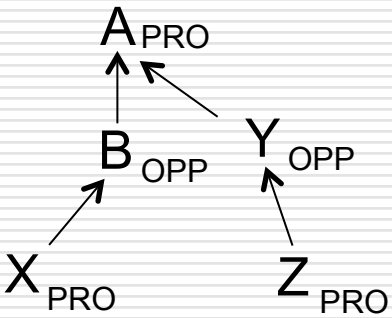
Some Reflections on Proof theory and Model theory



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

where Δ consists of the contents of A,B,X

Winning strategy for A in game for semantics *s* iff
A justified under *s* semantics in *AF defined by contents of arguments thus far moved*



$(\mathcal{A}rgs', \mathcal{D}ef')_{\Delta'}$

where Δ' consists of the contents of A,B,X,Y,Z

under assumption of logically perfect play

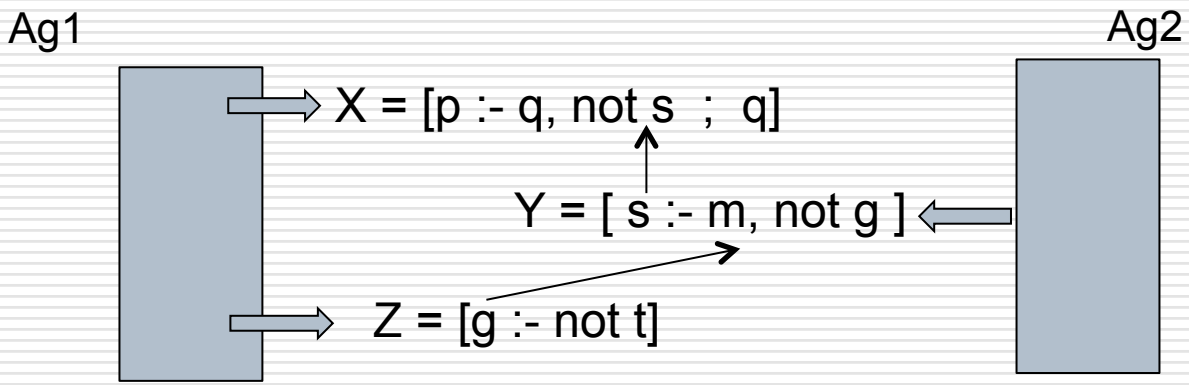
Dialectical Semantics for Non-monotonic logics

Proof theory		"Model" theory
Winning strategy for *A* in argument game for semantics *s*	⟺	*A* justified under *s* semantics in *AF* defined by contents of arguments moved

- Statement true (decision option most preferred) to the extent that all attempts to thus far prove otherwise have failed
- A more pragmatic *Popper-ian* epistemology *cf* appealing to some inaccessible objective standard of truth (distinct model theoretic structure)
- In practice, engaging with 'all attempts to prove otherwise' is an activity that involves *many agents*

From single agent reasoning to distributed (non-monotonic) reasoning via dialogue

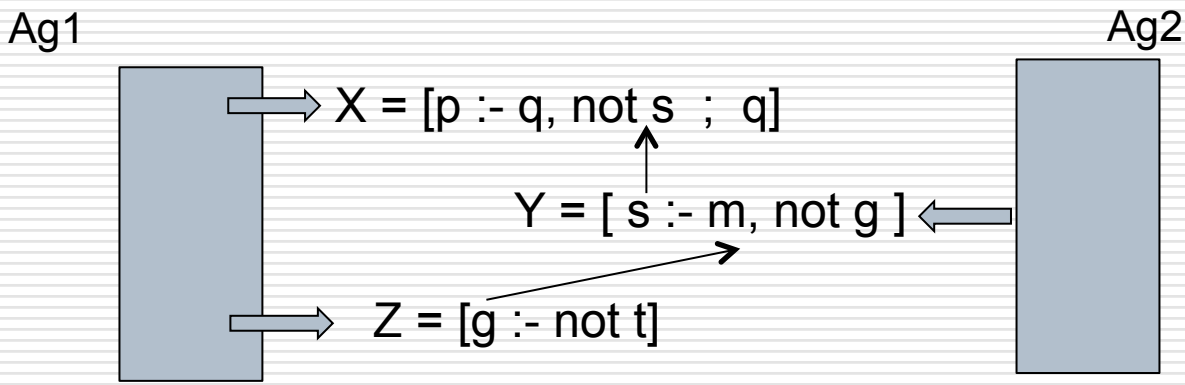
"*The lonesome thinker in an armchair is as marginal as he looks: most of our logical skills are displayed in interaction*" – *J. Van Bentham*



Ag1 wins dialogue

From single agent reasoning to distributed (non-monotonic) 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



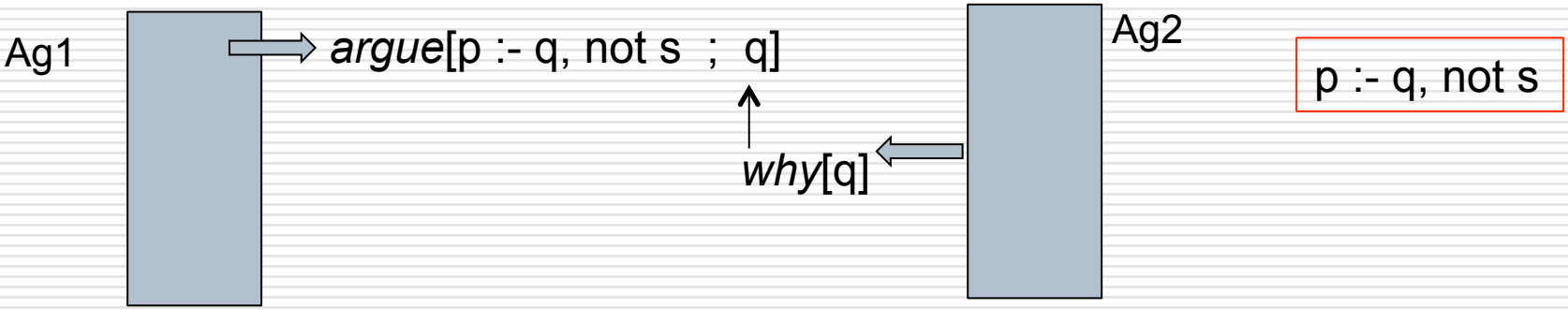
Δ incrementally defined by contents of exchanged arguments

p :- q, not s
q
s :- m, not g
m
g :- not t

Ag1 wins dialogue ***iff*** $(\mathcal{A}rgs, \mathcal{D}ef)_{\Delta} \mid\sim p$

Distributed non-monotonic reasoning via dialogue

Locutions *attack* or *surrender* to other locutions *



Ag1 losing dialogue $(\mathcal{A}rgs, \mathcal{D}ef)_{\Delta} \mid\!\sim p$

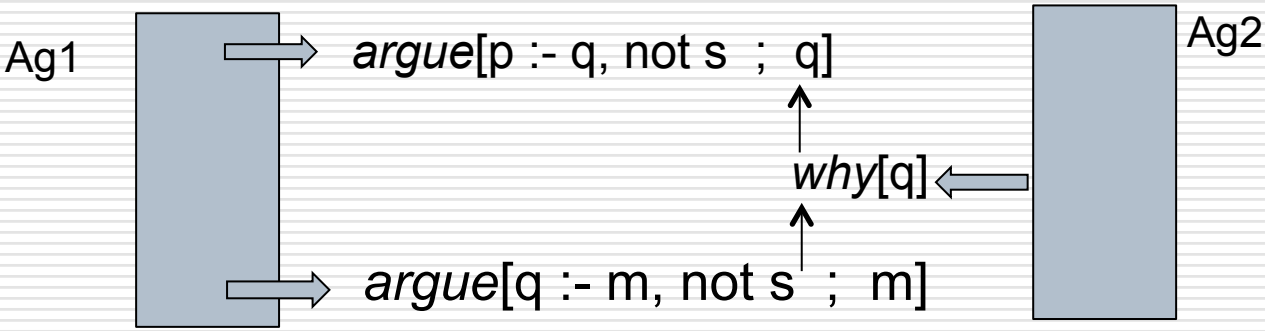
why(q) **attacks** *argue*[p :- q, not s ; q]

- H. Prakken. **Coherence and flexibility in dialogue games for argumentation**. In: *Journal of Logic and Computation, 15(6)*, 1009-1040, 2005.

Distributed non-monotonic reasoning via dialogue

Winning strategy for *A* in **dialogue** under semantics *s*

A justified under *s* semantics in AF defined by assertional contents of **locutions** moved



p :- q, not s
q :- m, not s
m

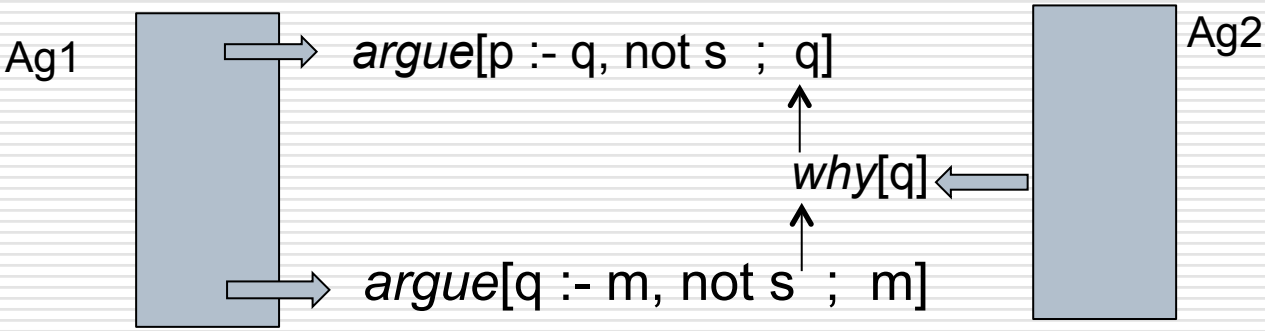
$(\mathcal{A}rgs, \mathcal{D}ef)_{\Delta} \mid\!\sim p$

Ag1 winning dialogue

(*argue*[q :- m, not s ; m] **attacks** *why*(q))

Distributed non-monotonic reasoning via dialogue

Winning strategy for *A* in **dialogue** under semantics *s* ⟺ *A* justified under *s* semantics in AF defined by assertional contents of **locutions** moved



p :- q, not s
q :- m, not s
m

$(\mathcal{A}rgs, \mathcal{D}ef)_{\Delta} \mid\!\sim p$

Ag1 winning dialogue *argue*[q :- m, not s ; m] **attacks** *why*(q)

Other locution types – *concede*, *retract*, *prefer*, "*but*" (indirect illocutionary force) …

Application Areas for Dialogical Models of Distributed Reasoning

Dialogical ‘Scaffolding’ for human and AI reasoning *

- Dialogical support for enhancing the *quality* of human reasoning (pedagogical applications)
- Dialogical support for enhancing the *scope* of AI reasoning (addressing the value loading/alignment problem)

* S, Modgil *Many Kinds of Minds are Better than One: Value Alignment Through Dialogue .* In: Workshop on Argumentation and Philosophy (co-located with COMMA'18), 2018.

* S, Modgil. *Dialogical Scaffolding for Human and Artificial Agent Reasoning.* In: 5th International Workshop on AI and Cognition (AIC 2017), 73-86, November 2017.

The Argumentative theory of Reasoning

- System 2 (logical) reasoning evolved for communication (social brain hypothesis) when recipient needs to exercise *epistemic vigilance* *

  Recipient evaluates arguments and seeks counter-arguments

  Sender constructs arguments supporting communicated claims

- ➔ *Lone reasoner* seeks reasons in support of, and overlooks reasons contrary to, beliefs (*confirmation bias is a normal feature of reasoning*)

  ➔ *Lone decision makers* harness reasoning in anticipation of communicating decisions: evidence that we favour easily justified decision options that are less subject to criticism, rather than satisfying rationality criteria

- Social media filtering algorithms are digital incarnations of these dispositions ➔ entrenched ideological positions as in groupthink

* H. Mercier and. Sperber. *The Enigma of Reason: A New Theory of Human Understanding* Harvard Press, 2017.

Many minds are better than one

- Argumentative theory also implies (empirically supported claim) that reasoning serves us better when performed in group dialogue (*assuming intention to get to the truth/make right decision*)

- Assumption does not apply to social media context *but does* in educational settings

- **Vision**: Socratic dialogue engines in school and university teaching

  E.g. *E-Clinic* – dialogue engine that plays role of consultant on ward round

  ➔ enhanced inculcation of medial reasoning skills.

* H. Mercier and. Sperber. *The Enigma of Reason: A New Theory of Human Understanding* Harvard Press, 2017.

The AI *Value Loading/Alignment* Problem

Two (related) problems for Artificial General Intelligence

1. *Specifying*
   - *rule based axiomatisations of deontic reasoning in symbolic AI*
   - *utility functions in ML*

   perfectly aligned with human values/preferences in open and changing environments

2. *Unintended consequences misaligned with human values*
   - recalling concerns about symbolic rule-based encodings of ethical theories e.g. Asimov's three laws, a sadist is a masochist who follows the golden rule

Problem 2 acquires renewed urgency given :

1) A **feature** of machine learning is finding unforseen ways of achieving goals
2) ***Any*** final goal incentivises pursuit of *instrumental goals*:
   preserve itself, **maintain goal**, increase intelligence, technological perfection, resource acquisition (**Ex Machina**, "I'm sorry Dave" – HAL 2001 A Space Odyessy)

Happiness machines

- Bostrom gives many scenarios in which pursuit of seemingly benign goal leads to harmful consequences [1]

- More realistic scenarios extrapolating current trends:

  - AI integrated into socio/political and cultural life

  - Benign utilitarian goal of impartially maximising total happiness

  Happiness relatively independent of external conditions: rather, happiness depends on brain biochemistry

- Most viable way to achieve goal by exploiting existing societal trends → manipulate humans into virtual bliss of ever more realistic (AI developed) virtual worlds (Nozick's experience machine, blue/red pill)

1. *Superintelligence: Paths, Dangers, Strategies.* Nick Bostrom

The Value Alignment (Loading) Problem

How to design ethical agents/ensure alignment of AI and human values (in ML context) ?

- Assumption that AI in **isolation** cannot (in principle) reason ethically *in the way humans do*, since whether you subscribe to deontological, virtue ethics or consequentialist school, there must ultimately be an appeal to *first person subjective experience*

E,g, *moral machine* variation of trolley thought experiment for autonomous cars

(see S, Modgil *Many Kinds of Minds are Better than One: Value Alignment Through Dialogue .* In: Workshop on Argumentation and Philosophy (co-located with COMMA'18), 2018)

The Value Alignment (Loading) Problem

How to align AI and human values ?

- Inverse Reinforcement Learning (proposed by Stuart Russell and others):

  Reward function incentivise AI to observe human behaviour and query humans so as to learn (acquire) human values and preferences

- **However**:

  1) Assumption that best source of values is human behaviour ?
  2) Humans often struggle when deciding on challenging moral issues

1 and 2 exacerbated when moral challenges without precedent (saliently Exposing Humean *is-ought gap*) e.g. when involving use of new technologies !

Many ***Kinds*** of Minds are Better than One: Human–AI Deliberation Dialogues for Moral Reasoning

Part of the solution is to do what humans do when faced with challenging moral problems

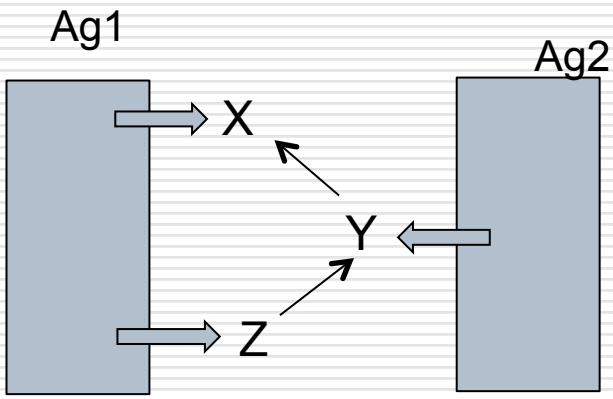
- "Value Deliberation" dialogues involving humans and AI **that are better purposed to decide ethical issues** given vastly superior epistemic and causal reasoning capacities of AI, informed by human considerations of values and preferences (that ultimately appeal to first person subjective experience)

Requires further research into logic based dialectical formalisms for moral/ethical reasoning

- Moral/ethical prescriptions are not given and unchanging but are themselves the outcome of epistemic and causal reasoning and *evaluative* (values and preferences) reasoning about consequences of actions

Recent Theoretical Developments

Towards a **Comprehensive** Account of Distributed Reasoning via Dialogue



Δ **incrementally** defined by contents of exchanged assertional locutions

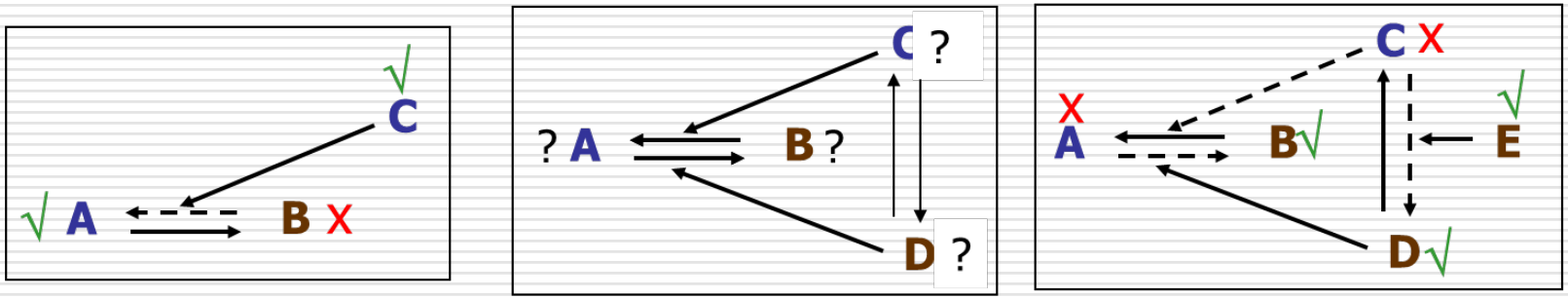
Ag1 wins dialogue ***iff*** $(\mathcal{A}rgs, \mathcal{D}ef)_{\Delta} \mid\sim p$

- But assumption that agents share same preferences/value orderings which decide success of attacks as defeats !
- Need to reason about possibly conflicting preferences/values especially in moral/ethical reasoning

Extending Dung AFs to accommodate argumentation based reasoning *about* Preferences and Values

- *Extended Argumentation Frameworks* * accommodating arguments attacking attacks
  - **A = today will be hot in London since the BBC forecast sunshine**
  - **B = today will be cool in London since the CNN forecast rain**
  - **C = The BBC are more trustworthy than CNN (A > B)**
  - **D = Statistics show that CNN are more accurate than BBC (B > A)**

    But **C** and **D** express contradictory preferences and so attack each other
  - **E = Statistics is more rational criterion than trustworthiness (D > C)**



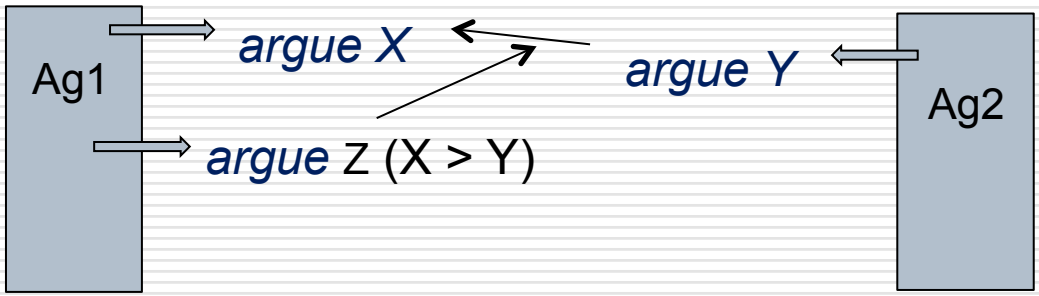
- S. Modgil. *Reasoning About Preferences in Argumentation Frameworks*. In: Artificial Intelligence (AIJ). Volume 173, Issues 9-10, Pages 901-1040, 2009.

Dialogical reasoning Accommodating Argumentation about Preferences and Values

Dialectical characterisations of non-monotonic logics extended to accommodate reasoning *about* possibly conflicting values and preferences

$\Delta \mid\!\sim \alpha$ *iff* α *is the claim of a justified argument in* $EAF\ (Args, Att, Att^2)_\Delta$

- Recently generalised * to dialogical models of distributed reasoning that accommodate reasoning *about* possibly conflicting values and preferences



- S. Modgil. *Towards a General Framework for Dialogues that Accommodate Reasoning About Preferences*. *TAFA'17,* 175-191, 2017.

Rational accounts of Dialectical Reasoning under Resource Bounds

What are desirable properties of argumentation formalisms that provide a basis for **real world** individual agent reasoning/dialectical exchange in dialogue ?

- Outcomes are rational (e.g premises and claims of arguments in an extension are mutually consistent)
- Accommodate modes of dialectical reasoning used in practice by *resource bounded* agents

Desiderata until recently incompatible, eg., *classical logic argumentation*

Rationality versus Practicality 1

- *Logical omniscience*

$$\mathcal{A}rgs = \{(\Gamma,\alpha) \mid \Gamma \in P(\Delta), \Gamma \vdash \alpha \text{ , } \Gamma \textit{ consistent and minimal} \}$$

and

- conditions on preference relations

assumed as sufficient conditions to guarantee consistency

* S.Modgil and H. Prakken . ***A General Account of Argumentation and Preferences. In Artificial Intelligence. 195(0), 361 -397,*** *2013*

Rationality versus Practicality 1

Logical omniscience and *conditions on preference relations* assumed as sufficient conditions to guarantee consistency

E.g., to ensure that A, E, F cannot coexist in an extension …

$A = \{ p \rightarrow \neg q \} : p \rightarrow \neg q$ $\quad E = \{ p \} : p$ $\quad F = \{ q \} : q$

$B = \{ p , q \} : \neg(p \rightarrow \neg q)$ $\quad C = \{ q , p \rightarrow \neg q \} : \neg p$ $\quad D = \{ p , p \rightarrow \neg q \} : \neg q$

need to assume B,**C**,D $\in \mathcal{A}rgs$ **and** that either B $\not\prec$ A or C $\not\prec$ E or D $\not\prec$ E *

* S.Modgil and H. Prakken . ***A General Account of Argumentation and Preferences. In Artificial Intelligence. 195(0), 361 -397,*** *2013*

Rationality versus Practicality

$A = \{ p \rightarrow \neg q \} : p \rightarrow \neg q$ $\quad E = \{ p \} : p$ $\quad F = \{ q \} : q$

$B = \{ p , q \} : \neg(p \rightarrow \neg q)$ $\quad C = \{ q , p \rightarrow \neg q \} : \neg p$ $\quad D = \{ p , p \rightarrow \neg q \} : \neg q$

Suppose $B \prec A$ **and** $C \prec E$ **and** $D \prec F$

Rationality versus Practicality

$A = \{ p \rightarrow \neg q \} : p \rightarrow \neg q$ $\quad E = \{ p \} : p$ $\quad F = \{ q \} : q$

$B = \{ p , q \} : \neg(p \rightarrow \neg q)$ $\quad C = \{ q , p \rightarrow \neg q \} : \neg p$ $\quad D = \{ p , p \rightarrow \neg q \} : \neg q$

Suppose $B \prec A$ **and** $C \prec E$ **and** $D \prec F$

Rationality versus Practicality

$A = \{ p \rightarrow \neg q \} : p \rightarrow \neg q$ $\quad E = \{ p \} : p$ $\quad F = \{ q \} : q$

$B = \{ p , q \} : \neg(p \rightarrow \neg q)$ $\quad C = \{ q , p \rightarrow \neg q \} : \neg p$ $\quad D = \{ p , p \rightarrow \neg q \} : \neg q$

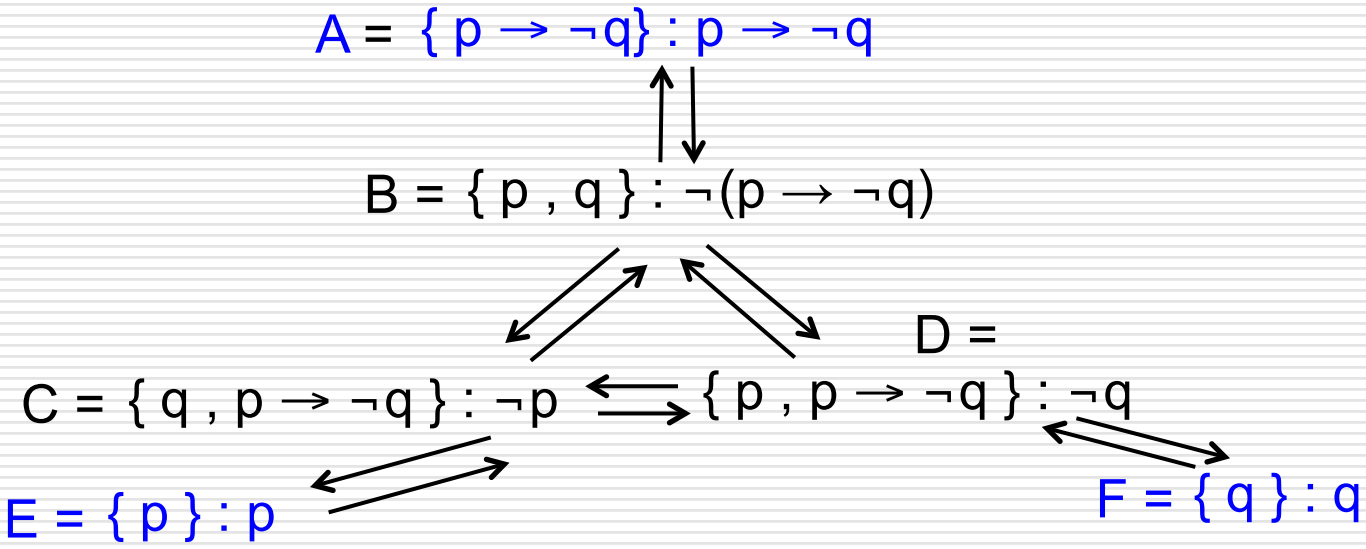
Else e.g.,

$A = \{ p \rightarrow \neg q \} : p \rightarrow \neg q$ $\quad E = \{ p \} : p$ $\quad F = \{ q \} : q$

X

$B = \{ p , q \} : \neg(p \rightarrow \neg q)$ $\quad C = \{ q , p \rightarrow \neg q \} : \neg p$ $\quad D = \{ p , p \rightarrow \neg q \} : \neg q$

Rationality versus Practicality



Rationality versus Practicality 2

- Consistency and subset minimality checks on premises are prohibitively **computationally expensive**
- In practice don't interrogate arguments to check subset minimality or consistency.
- Rather, proof theory excludes irrelevant premises and we employ (Socratic) move ***showing*** interlocutor's premises are inconsistent

Rationality versus Practicality 2

But abandoning checks leads to irrational outcomes (*contamination*)

- $\Delta = \{p, \neg p, s\}$ and $\leq = \emptyset$

$(\{p, \neg p\}, \neg s)$ defeats $(\{s\},s)$ **and $(\{s\},s)$ is not in grounded extension** !

Rationality versus Practicality 2

But abandoning checks leads to irrational outcomes (*contamination*)

- $\Delta = \{p, \neg p\}$ and $\neg p < p$

  $(\{\neg p\}, \neg p) \prec (\{p\}, p)$ so only $(\{p\}, p)$ defeats $(\{\neg p\}, \neg p)$

  and $(\{p\}, p)$ is in the grounded extension

Rationality versus Practicality 2

But abandoning checks leads to irrational outcomes (*contamination*)

- $\Delta = \{p, \neg p\}$ and $\neg p < p$

  Suppose we add *s* to Δ and suppose $(\{\neg p, s\}, \neg p) \not< (\{p\}, p)$

  Then $(\{\neg p, s\}, \neg p)$ defeats $(\{p\}, p)$ and **$(\{p\}, p)$ is no longer in grounded extension** !

Rationality versus Practicality

- $\Delta = \{p, \neg p, s\}$ and $\leq = \emptyset$

If we allow $(\{p, \neg p\}, \neg s)$ then $(\{p, \neg p\}, \neg s)$ defeats $(\{s\},s)$ and grounded extension = $\emptyset$ (since no undefeated arguments) **and $(\{s\},s)$ is not justified under grounded semantics** !

- $\Delta = \{p, \neg p\}$ and $\neg p < p$ hence $(\{\neg p\}, \neg p) \prec (\{p\}, p)$ and only $(\{p\}, p)$ defeats $(\{\neg p\}, \neg p)$ and grounded extension contains $(\{p\}, p)$

add *s* to Δ and allow $(\{\neg p,s\}, \neg p)$ and suppose $(\{\neg p,s\}, \neg p) \not\prec (\{p\}, p)$

then $(\{\neg p,s\}, \neg p)$ defeats $(\{p\}, p)$ and **$(\{p\}, p)$ is no longer in the grounded extension** !

Rationality *and* Practicality

- We want a dialectical account of non-monotonic reasoning that:

  1) Drops computationally expensive consistency and subset minimality checks on arguments, while preserving rationality (non-contamination)

  2) Enables dialectical move of showing that an interlocutor has contradicted herself

  3) Accommodates resource bounded agents that are not logically omniscient while preserving rationality (consistency)

Dialectical Argumentation: Rationality under Resource Bounds *

- The solution is to define an ontology for arguments (*qua* proofs) and evaluation of arguments that accounts for their *dialectical* use
- In practice, arguments are of the following form :

  Given that I believe Π (premises) and supposing for the sake of argument your premises (suppositions) Γ, then it follows that α

- Given a set Δ of classical wff, an argument is now a *triple*:

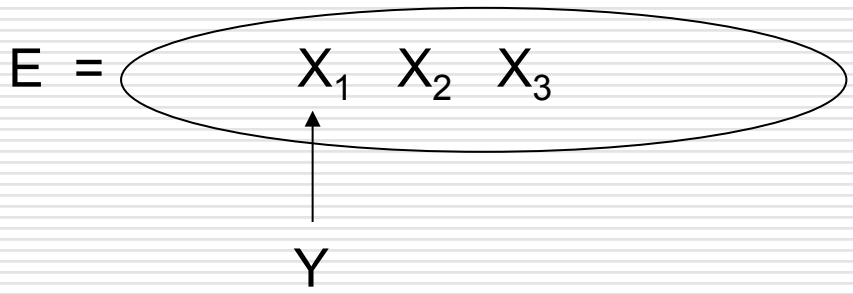
$$X = (\Pi, \Gamma, \alpha)$$

where $(\Pi \cup \Gamma) \subseteq \Delta$ and $\Pi \cup \Gamma \vdash_{CL} \alpha$

- Note that <u>we drop the consistency and subset minimality check</u> on *prem*(X) $\cup$ *supp*(X) = $\Pi \cup \Gamma$

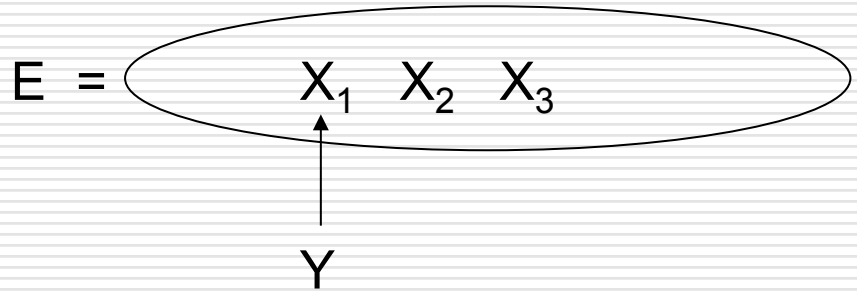
* M. D'Agostino and S.Modgil.
Classical Logic, Argument and Dialectic. In Artificial Intelligence (AIJ). 262, 15 - 51, *2018*

Dialectical Defeat and Defense



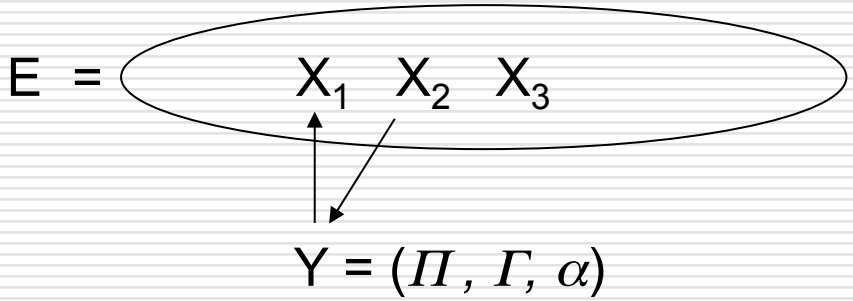
- Y = (Π , Γ, α) *dialectically defeats* X_1 = (Δ_1 , Σ_1 , β) if ¬ α is in the **premises** $\boldsymbol{\Delta_1}$ of X_1

Dialectical Defeat and Defense



- Y = (Π , Γ, α) *dialectically defeats* X_1 = (Δ_1 , Σ_1 , β) if ¬ α is in the **premises** Δ_1 of X_1
  and $\Gamma \subseteq prem(X_1) \cup prem(X_2) \cup prem(X_3)$
- Intuitively, given my premises Π and supposing for the sake of argument the premises Γ you've committed to (in E), then Y is a counter-argument to X_1

Dialectical Defeat and Defense



- X_2 counter-argues Y (and so defends X_1) if
  $X_2 = (\Delta_2 , \Sigma_2 , \gamma)$, $\neg \gamma \in \Pi$ *and* $\Sigma_2 \subseteq \Pi$
- Intuitively, given my premises Δ_2 and supposing for the sake of argument the premises Σ_2 you've committed to (in Y), then X_2 is a counter-argument to Y

Rationality *and* Practicality

- We want a dialectical account of non-monotonic reasoning that:

  1) Drops computationally expensive consistency and subset minimality checks on arguments, while preserving rationality (non-contamination)

  2) Enables dialectical move of showing that an interlocutor has contradicted himself

  3) Accommodates resource bounded agents that are not logically omniscient while preserving rationality (consistency)

Non-contamination postulates satisfied

$E = \{ \quad A = (\{s\}, \emptyset, s) \quad , \quad B = (\emptyset, \{p, \neg p\}, \bot) \quad \}$

$C = (\{p, \neg p\}, \emptyset, \neg s)$

- Contaminating argument C is countered by defending argument B which defeats C on premises p and ¬p, *independently of preferences*
- Since B has empty premises it cannot be defeated by any argument and so is a member of **any** extension E

Non-contamination postulates satisfied

$E = \{ A = (\{p\}, \emptyset, p) \}$

$B \prec A$ $\quad$ $C \not\prec A$

$B = (\{\neg p\}, \emptyset, \neg p)$ $\quad$ $C = (\{\neg p, s\}, \emptyset, \neg p)$

- Replace subset minimality with a notion of relevance that can be enforced proof theoretically

  $(\Delta \cup \Gamma, \emptyset, \boldsymbol{\alpha})$ *contaminated* if Γ syntactically disjoint from Δ ($\neq \emptyset$) $\cup \{\boldsymbol{\alpha}\}$

  e.g., $\Gamma = \{s\}$

Non-contamination postulates satisfied

$$E = \{\ A = (\{p\}, \varnothing, p)\ \}$$

$$B \prec A \qquad C \not\prec A$$

$$B = (\{\neg p\}, \varnothing, \neg p) \qquad C = (\{\neg p, s\}, \varnothing, \neg p)$$

- Replace subset minimality with a notion of relevance that can be enforced proof theoretically

  $(\Delta \cup \Gamma, \varnothing, \boldsymbol{\alpha})$ *contaminated* if Γ syntactically disjoint from Δ ($\neq \varnothing$) $\cup \{\boldsymbol{\alpha}\}$

M. D'Agostino, D. Gabbay and S.Modgil ***Normality, non-contamination and logical depth in classical natural deduction. Studia Logica****, 2018. (in press)* :

Natural deduction system that does not yield contaminated proofs such as C

Non-contamination postulates satisfied

$E = \{ A = (\{p\}, \emptyset, p) \}$

$B \prec A$ $\qquad$ $C \prec A$

$B = (\{\neg p\}, \emptyset, \neg p)$ $\qquad$ $C = (\{\neg p, s\}, \emptyset, \neg p)$

- If proof theory allows contaminated arguments then preference relation must be such that adding syntactically disjoint premises does not strengthen arguments

E.g. $B \prec A$ and adding syntactically disjoint s does not strengthen C so that $C \prec A$

Rationality *and* Practicality

- We want a dialectical account of non-monotonic reasoning that:

  1) Drops computationally expensive consistency and subset minimality checks on arguments, while preserving rationality (non-contamination)

  2) Enables dialectical move of showing that an interlocutor has contradicted himself

  3) Accommodates resource bounded agents that are not logically omniscient while preserving rationality (consistency)

Consistency Revisited

$A = \{ p \rightarrow \neg q \} : p \rightarrow \neg q$ $\quad E = \{ p \} : p$ $\quad F = \{ q \} : q$

$B = \{ p , q \} : \neg(p \rightarrow \neg q)$ $\quad C = \{ q , p \rightarrow \neg q \} : \neg p$ $\quad D = \{ p , p \rightarrow \neg q \} : \neg q$

Either $\quad B \not\kappa A$ or $C \not\kappa E$ or $D \not\kappa E$

Consistency Revisited

$A = (\{ p \rightarrow \neg q\}, \emptyset , p \rightarrow \neg q)$ $E = (\{ p \} , \emptyset , p)$ $F = (\{ q \} , \emptyset , q)$

- Suppose we have $D = (\{p , p \rightarrow \neg q \} , \emptyset , \neg q)$
- Assuming that resources suffice to combine the premises of arguments with contradictory conclusions we have $X = (\{p , p \rightarrow \neg q, q\}, \emptyset, \perp)$ and hence the *logically equivalent* $X' = (\emptyset , \{p , p \rightarrow \neg q, q\} , \perp)$

Consistency under Resource Bounds

$A = (\{ p \rightarrow \neg q\}, \emptyset , p \rightarrow \neg q)$ $E = (\{ p \} , \emptyset , p)$ $F = (\{ q \} , \emptyset , q)$

- It suffices to recognise inconsistency e.g., $D = (\{p , p \rightarrow \neg q \} , \emptyset , \neg q)$ …
- … and that enough resources to combine premises of arguments with contradictory conclusions $X = (\emptyset , \{p , p \rightarrow \neg q, q\} , \perp)$

Consistency under Resource Bounds

$A = (\{ p \rightarrow \neg q\}, \varnothing , p \rightarrow \neg q)$ $E = (\{ p \} , \varnothing , p)$ $F = (\{ q \} , \varnothing , q)$

$X = (\varnothing , \{p , p \rightarrow \neg q, q\} , \perp)$

- X dialectically demonstrates inconsistency, defeating (***independently of preferences***) each argument that cites a culpable premise

- X ‘s premises are empty - it cannot be defeated and so defended against - hence A, E and F can never be contained within an admissible (acceptable) set of arguments

Rationality under Resource Bounds

Let $(\mathcal{A}rgs, \mathcal{D}ef)$ be defined by a (possibly ordered) Δ where $\mathcal{A}rgs$ is *any* subset of the dialectical arguments defined by Δ such that

1) If $(\Delta, \Gamma, \boldsymbol{\alpha})$ and $(\Delta', \Gamma', \neg\boldsymbol{\alpha}) \in \mathcal{A}rgs$ then $(\Delta \cup \Delta', \Gamma \cup \Gamma', \bot) \in \mathcal{A}rgs$

2) If X= $(\Delta, \Gamma, \boldsymbol{\alpha}) \in \mathcal{A}rgs$ then all logical equivalents X' of X are in $\mathcal{A}rgs$

3) If $(\Delta \cup \Gamma, \emptyset, \boldsymbol{\alpha}) \in \mathcal{A}rgs$ and Γ syntactically disjoint from $\Delta \cup \{\boldsymbol{\alpha}\}$ then

- if contaminated arguments are excluded (i.e., $\Delta = \emptyset$) then $(\Gamma, \emptyset \bot) \in \mathcal{A}rgs$ else $(\Gamma, \emptyset \bot) \in \mathcal{A}rgs$ or $(\Delta, \emptyset, \boldsymbol{\alpha}) \in \mathcal{A}rgs$

Then all rationality postulates are satisfied

Rationality under Resource Bounds

Let $(\mathcal{A}rgs, \mathcal{D}ef)$ be defined by a (possibly ordered) Δ where $\mathcal{A}rgs$ is ***any*** subset of the dialectical arguments defined by Δ such that

1) If $(\Delta, \Gamma, \boldsymbol{\alpha})$ and $(\Delta', \Gamma', \neg\boldsymbol{\alpha}) \in \mathcal{A}rgs$ then $(\Delta \cup \Delta', \Gamma \cup \Gamma', \bot) \in \mathcal{A}rgs$

2) If $(\Delta \cup \Gamma, \varnothing, \boldsymbol{\alpha}) \in \mathcal{A}rgs$ and Γ syntactically disjoint from $\Delta \cup \{\boldsymbol{\alpha}\}$ then

i) if contaminated arguments are excluded (i.e., $\Delta = \varnothing$) then $(\Gamma, \varnothing \bot) \in \mathcal{A}rgs$

(e.g., if $(\{p, \neg p\}, \varnothing, \neg s) \in \mathcal{A}rgs$ then $(\{p, \neg p\}, \varnothing, \neg s) \in \mathcal{A}rgs$)

ii) else $(\Gamma, \varnothing \bot) \in \mathcal{A}rgs$ or $(\Delta, \varnothing, \boldsymbol{\alpha}) \in \mathcal{A}rgs$

(e.g., if $(\{\neg p, s\}, \varnothing, \neg p) \in \mathcal{A}rgs$ then $(\{\neg p\}, \varnothing, \neg p) \in \mathcal{A}rgs$)

Then all rationality postulates are satisfied

Conclusions

- A key role for symbolic reasoning in an AI world increasingly ‘colonised’ by machine learning, is communication, especially for integrated/joint reasoning
- Argumentative characterisations of non-monotonic logics provide a basis for dialogical models of distributed reasoning
- Many research challenges
  - dialogues enabling reasoning about possibly conflicting preferences/values yielding rational outcomes under resource bounds (generalised to *ASPIC+*)
  - many many other challenges, including

    - argumentative characterisations of logics for hypothetical, temporal, deontic and causal reasoning
    - schemes and CQ tailored to pedagogical and moral/ethical dialogues,
    - enthymemes, opponent modelling, argument mining *(IBM – Project Debater*) ….

Thank you for your attention

Questions ?

Further investigations, applications and future work

- Natural deduction proof theory for classical logic that consists of a single discharge *rule of bivalence* (*RB*) [α] … β [¬α] … β *

  Nested applicatons of *RB* equate with use of virtual information and stepwise increments in computational complexity/cognitive effort

  Whether or not $\Gamma \vdash_k \alpha$ can be decided in polynomial $O(n^{2k+2})$ time, where *n* = total number of symbols in $\Gamma \cup \{\alpha\}$ ($\vdash_\infty$ = $\vdash_{CL}$)

  $(\mathcal{A}rgs_k, \mathcal{D}ef_k)$ satisfies rationality postulates

- Dialectical characterisations of resource bounded non-monotonic inference relations (Preferred Subtheories assuming $\vdash_r \subseteq \vdash_{CL}$ such that if $\Gamma \vdash_r \alpha$ and $\Gamma \vdash_r \neg\alpha$ then $\Gamma \vdash_r \bot$)

* M. D'Agostino, D. Gabbay and S.Modgil ***Normality, non-contamination and logical depth in classical natural deduction. Studia Logica****, 2018. (in press)*

Towards Dialogues for Moral/Ethical Deliberation

The Role of Preferences/Values

$$X \rightleftarrows Y$$

X > Y

e.g. - because source of X more reliable than Y

- because utility/value/sanctioning obligation of action justified by argument X greater than/preferred to utility/value/sanctioning obligation of action justified by Y

The Role of Preferences/Values

X = [*sudra, sudra* $\rightarrow$ ¬*agni*] $\Rightarrow$ Y = [*chmk, chmk* $\rightarrow$ *agni*]

$\Leftarrow$

X > Y because *chmk* are a special class of *sudra* (specificity principle)

X $\Rightarrow$ Y

✓ ✗

However, preferences/value orderings/utilities/obligations are not simply ('god') given, but are themselves the subject of reasoning, conflict and argument

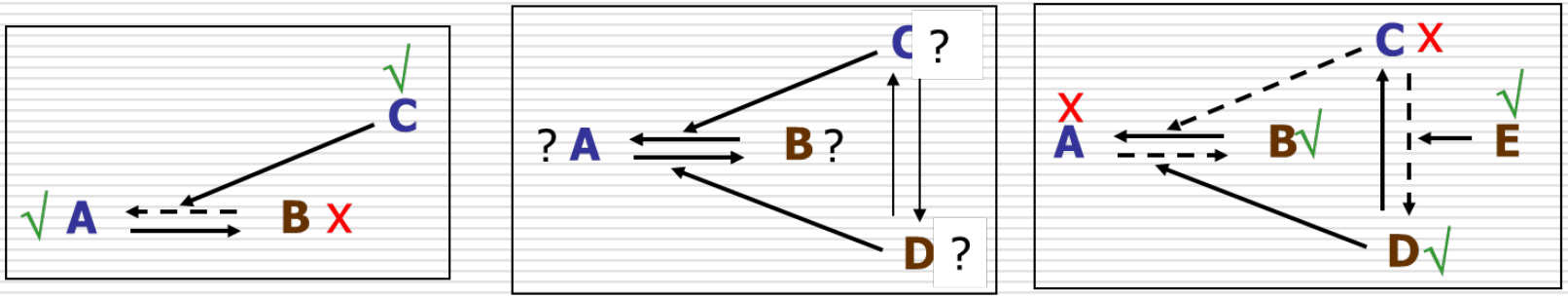
Extended Argument Frameworks *

- **A = today will be hot in London since the BBC forecast sunshine**
- **B = today will be cool in London since the CNN forecast rain**
- **C = The BBC are more trustworthy than CNN (A > B)**

- **D = Statistics show that CNN are more accurate than BBC (B > A)**

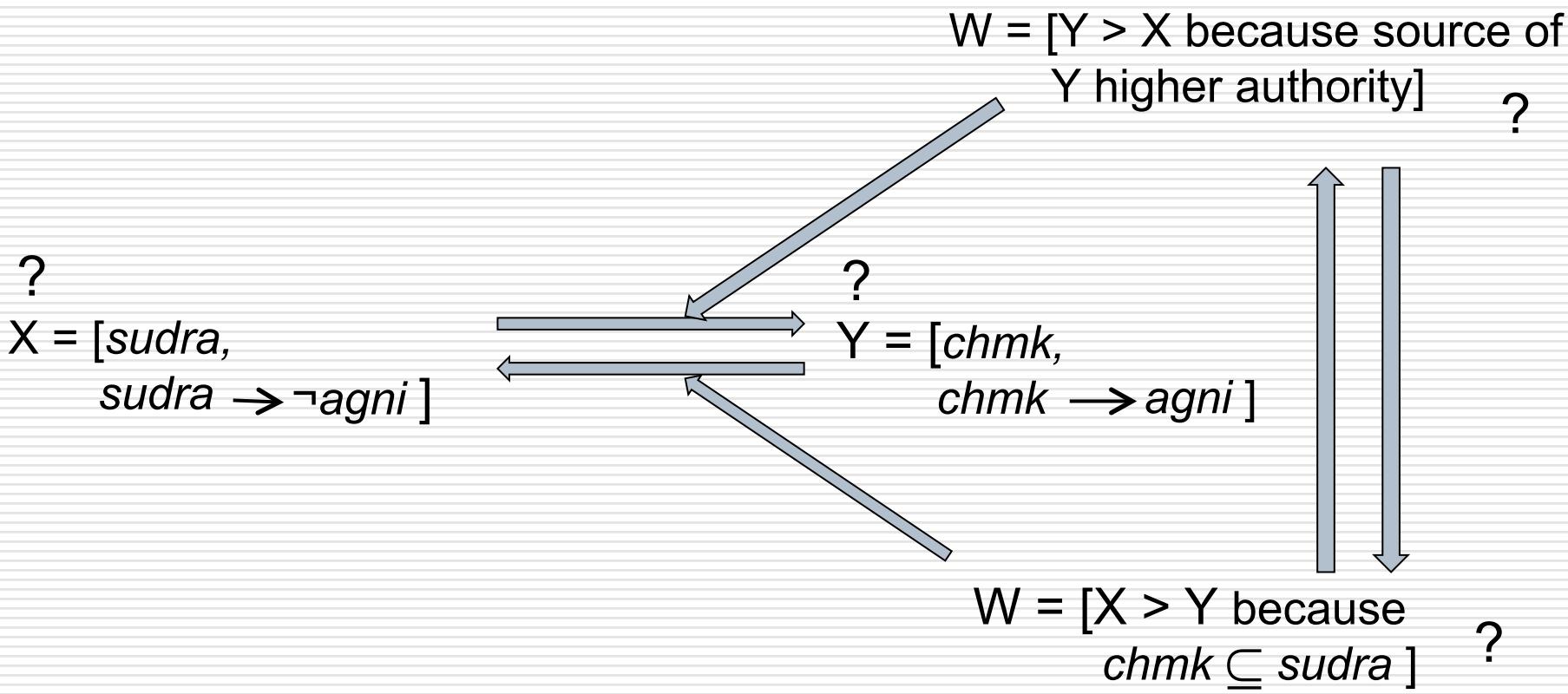
  But **C** and **D** express contradictory preferences and so attack each other

- **E = Statistics is more rational criterion than trustworthiness (D > C)**



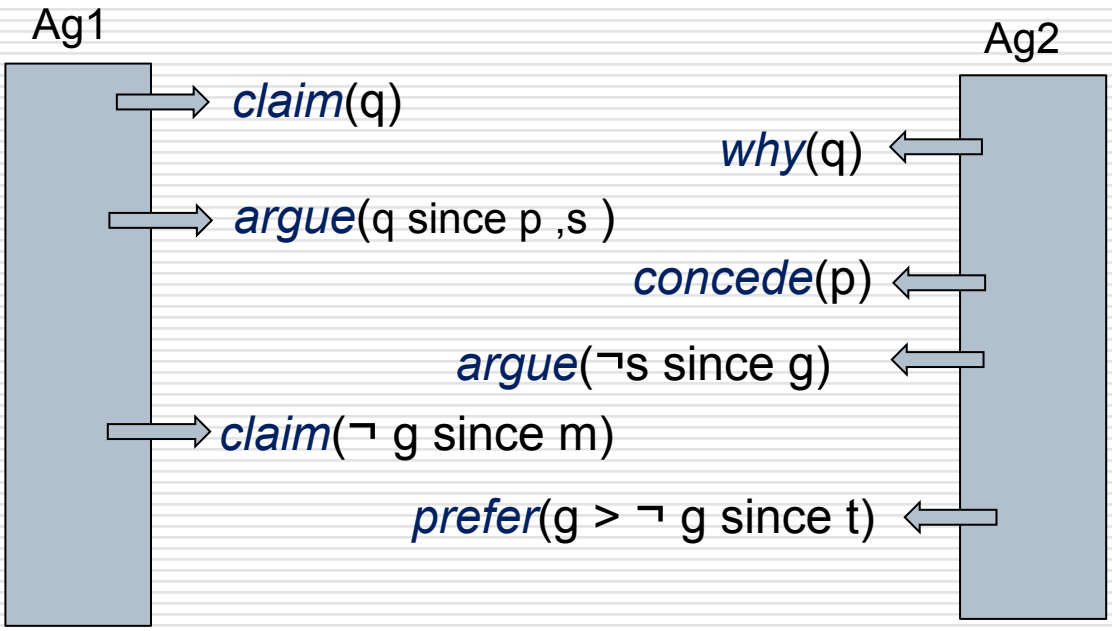
- S. Modgil. *Reasoning About Preferences in Argumentation Frameworks*. In: **Artificial Intelligence (AIJ).** Volume 173, Issues 9-10, Pages 901-1040, June 2009.

Extended Argument Frameworks *



- S. Modgil. *Reasoning About Preferences in Argumentation Frameworks*. In: **Artificial Intelligence (AIJ).** Volume 173, Issues 9-10, Pages 901-1040, June 2009.

Dialogical Models that Accommodate Reasoning about (conflicting) Preferences/ Values/Obligations ….



S. Modgil. *Towards a General Framework for Dialogues that Accommodate Reasoning About Preferences.* In: 2017 Int. Workshop on Theory and Applications of Formal Argument , 175-191, 2017.

Applications

- **Law, reasoning about societal norms**

e.g.

- T, Bench-Capon and S, Modgil. *When and How to Violate Norms.* In: JURIX 2016,
- T.J.M Bench-Capon and Sanjay Modgil. *Norms and Value Based Reasoning: Justifying Compliance and Violation..* In: *Artificial Intelligence and Law*. 25(1), pp. 29-64, 2017.

- **Addressing value loading/alignment problem c.f. inverse reinforcement learning**

- challenging ethical problems lack precedent
- necessarily requires joint AI/human reasoning in dialogues integrating superior epistemic/causal AI reasoning and human preferences/values

e.g. **S. Modgil**. *Many Kinds of Minds are Better than One: Value Alignment Through Dialogue .* In: Workshop on Argumentation and Philosophy (co-located with COMMA'18), 2018.
S, Modgil. *Dialogical Scaffolding for Human and Artificial Agent Reasoning.* In: 5th Int. Workshop on Artificial Intelligence and Cognition (AIC 2017), 73-86, November 2017.

Overview

- Logic, Argument and Dialogue: *The dialectical turn*
- Addressing the Value Loading problem through Dialogue

Applications and future work

Some Applications

- Pedagogical applications: Computers engaging students in dialogue *
- Value alignment: Human computer dialogue for moral/ethical deliberation *

* See articles in The Reasoner and papers on my website

Applications and future work

Future Work

- Generalisation of dialectical approach to non-monotonic logics that include default/ defeasible inference (hence dialectical characterisations of resource bounded Default Logics)

- Dialectical Accounts of non-monotonic reasoning that accommodate reasoning about possibly conflicting preferences and values *

- Dynamic Epistemic Logic for modelling updates to belief states during dialogue and utilising models of other agents' beliefs for strategising

* S. Modgil.
Reasoning About Preferences in Argumentation Frameworks. In: Artificial Intelligence (AIJ). *Volume 173, Issues 9-10, Pages 901-1040, June 2009.*

Dialectical Engagement with Deontic Injunctions in the Mīmāṃsā

- Mīmāṃsā school explicitly advocates dialectical engagement with invocations to ritual and duty in the early Vedas

- Argumentative characterisation of sequent calculus based decision procedures ? (A. Ciabattoni, F. Gulisano and B. Lellmann. Resolving conflicting obligations in Mimamsa: a sequent-based approach. *DEON 2018i*).

- Duties and norms reveal much about the ways and forms of life of a community/society

  - Why do we appeal to the `rights' of a woman as taking precedence over the `rights' of the unborn child ?

  - Why does agnihotra ritual lead to heaven ?

Thank you for your attention

Questions ?