Dialogical Scaffolding for Human and Artificial Agent Reasoning

Marcello D'Agostino
Department of Philosophy
University of Milan
marcello.dagostino@unimi.it

Sanjay Modgil
Department of Informatics
King's College London
sanjay.modgil@kcl.ac.uk

Overview

- From argumentative formalisations of nm reasoning to dialogue
- Applications of argumentation based dialogue
- A dialectical account of argumentation – towards fully rational accounts of non-monotonic reasoning under resource bounds

Resurrecting Dialogical Conceptions of Logic

- ❑ Early dialectical/dialogical conceptions of logic (from the Greeks onwards) supplanted by more solipsistic emphasis on individual agents reasoning using logic

- ❑ Lorenzen and Lorenz, Keith Stenning, Johan van Benthem, Catarina D. Novaes… rehabilitating dialectical/dynamic accounts of (typically deductive monotonic) logics

- ❑ However logical reasoning in the form of **adversarial communication as** witnessed in practice, in debate, moral reasoning, scientific enquiry etc – focus on arbitrating amongst decision options and contentious/conflicting beliefs ➔

- ❑ Dialogical formalisations of non-monotonic logics that supplement deductive logics with defeasible inference

Argumentative Formalisations of Non-monotonic Reasoning

Belief Base + Deductive and/or Defeasible Inference Rules + Preference Information

$\mid\!\sim_{\text{NmL}} \quad \alpha$

NmL = *Preferred Subtheories*
Prioritised Default Logic,
Defeasible Logic, Logic Programming

Argumentative Formalisations of Non-monotonic Reasoning

Belief Base + Deductive and/or Defeasible Inference Rules + Preference Information

$\mid\!\sim_{NmL} \quad \alpha$

NmL = Preferred Subtheories
Prioritised Default Logic,
Defeasible Logic, Logic Programming ….

Args

e.g., A = $(\{a, a \longrightarrow b\}, b)$

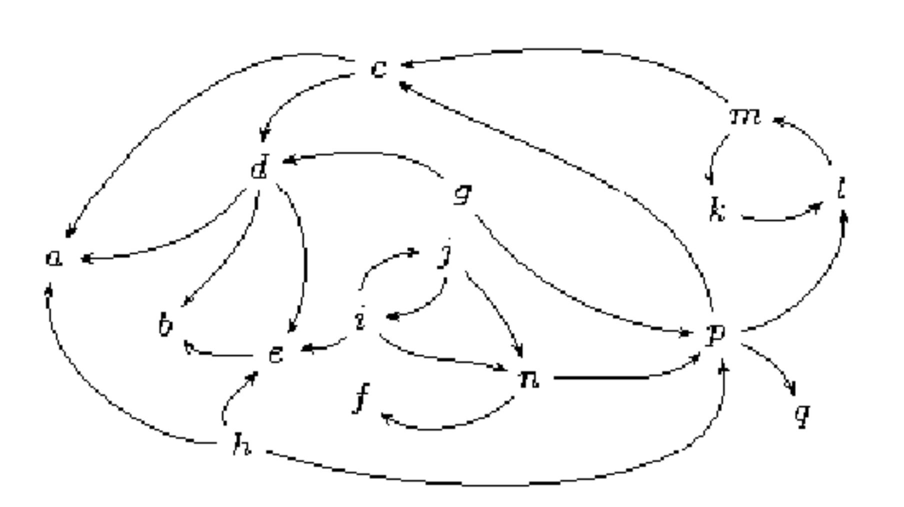
Defeat

e.g. A defeats B = $(\{\neg b\}, \neg b)$

Argumentative Formalisations of Non-monotonic Reasoning

Belief Base + Deductive and/or Defeasible Inference Rules
+ Preference Informmation

Argument Framework
< $\mathcal{Args}$, $\mathcal{Defeats}$ >



$|\!\sim_{\text{NmL}} \quad \alpha$

NmL = Preferred Subtheories
Prioritised Default Logic,
Defeasible Logic, Logic Programming

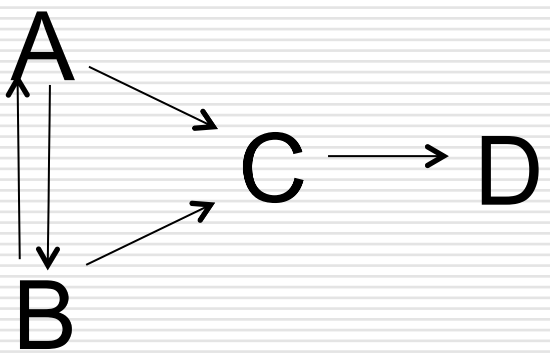
iff

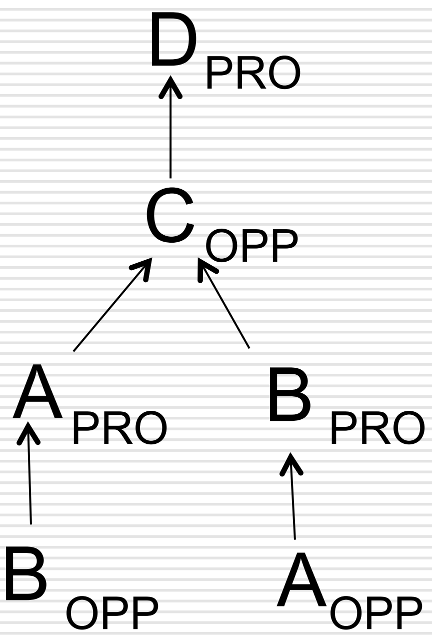
α is the claim of a winning (justified) argument

[1] P. M. Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. Artificial Intelligence, 77:321–357, 1995.

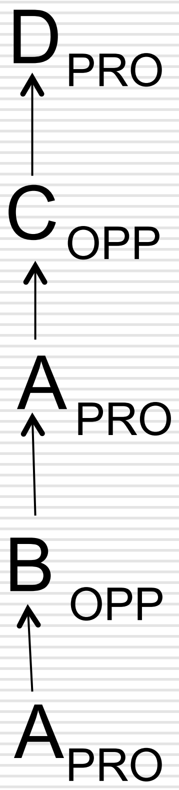
Argument Game Proof Theories

- Argument game proof theories – PRO v OPP – establish whether argument in a framework justified under a given *semantics* (i.e., burden of proof) – equivalently whether claim is *nm* inference from underlying belief base





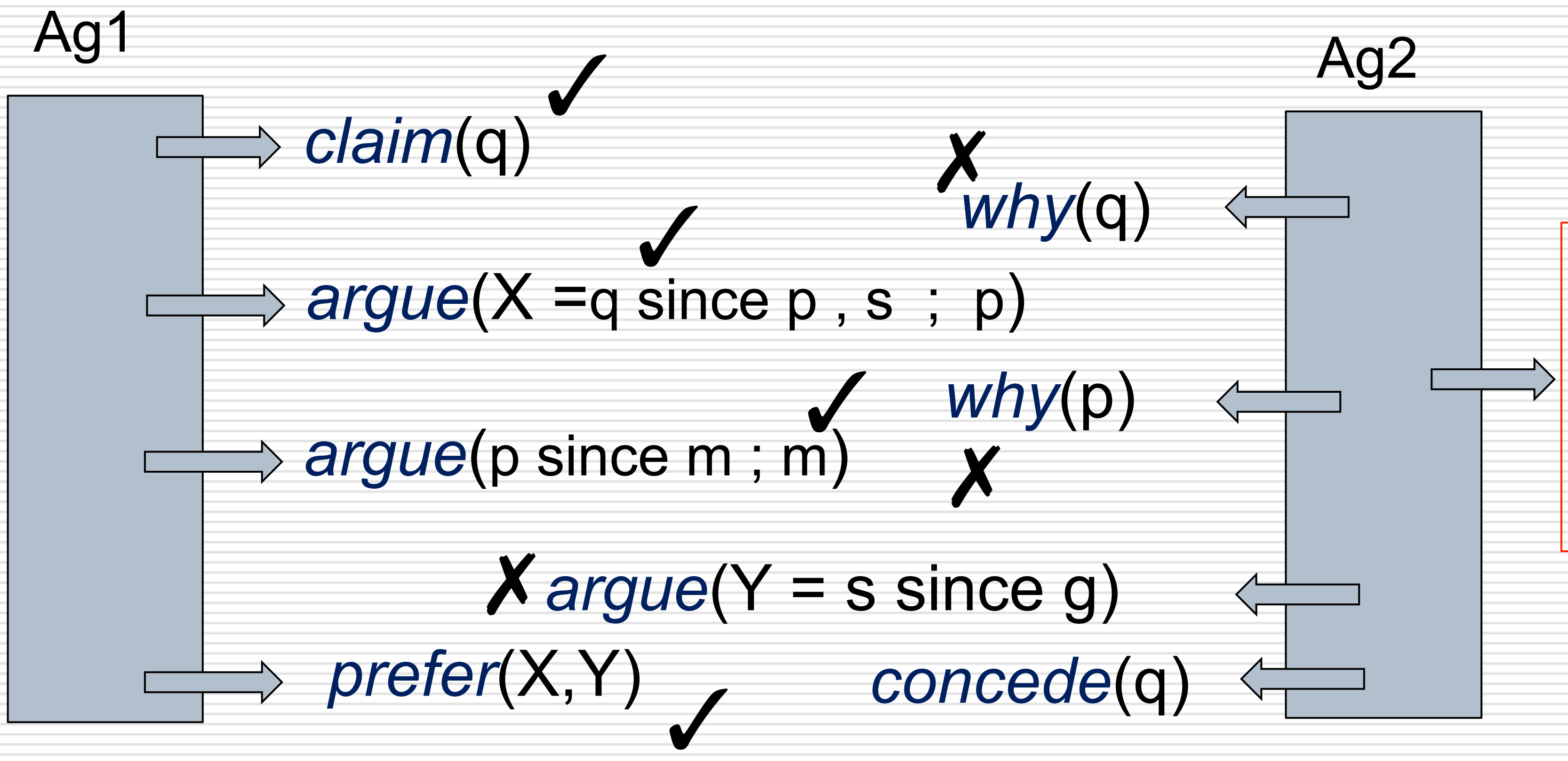
PRO loses game (D is not justified under **grounded** semantics)



PRO wins game – D justified under **preferred** semantics

S Modgil, M Caminada. *Proof theories and algorithms for abstract argumentation frameworks*. Argumentation in artificial intelligence, 105-129, 2009

Generalising Argument Games: From single agent reasoning to distributed non-monotonic reasoning via dialogue



Ag1

Ag2

claim(q) ✓

✗ *why*(q)

✓

argue(X =q since p , s ; p)

✓ *why*(p)

argue(p since m ; m) ✗

✗ *argue*(Y = s since g)

prefer(X,Y) ✓ *concede*(q)

Δ

Belief base **incrementally** defined by declarative contents of locutions

Dialectical status of locution (claim α) is winning

iff $\Delta \mid\!\sim \alpha$

H. Prakken. *Coherence and flexibility in dialogue games for argumentation*. Journal of logic and computation 2005

S Modgil. *Towards a general framework for dialogues that accommodate reasoning about preferences* Int. Workshop on Theories and Applications of Formal Argumentation TAFA 2017.

- **Dialogical support for:**

  1) Enhancing quality and scope of human reasoning;

  2) Enabling joint human and AI reasoning

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).

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

Sperber and Mercier's 'argumentative theory of reasoning' [1,2]

- ❑ Social role of dialogical models of nm reasoning supported by argumentative theory of reasoning

- ❑ Reasoning evolved to *asymmetrically* produce and evaluate arguments when communicating

Explains why reasoning alone leads us astray:

- confirmation bias
- reasoning drives people to decisions for which they can find arguments, rather than decisions that are optimal

- ❑ Theory also explains why reasoning through dialogue leads to better beliefs/decisions

1. H. Mercier and D. Sperber. *Why do humans reason? arguments for an argumentative theory*. Behavioral and Brain Sciences, 34(2):57–747, 2011.
2. H. Mercier and D. Sperber. *The Enigma of Reason: A New Theory of Human Understanding*, 2017

Applications of Argumentation-based Dialogue for Scaffolding Human Reasoning

- Normative dialectical guidance for human-human dialogue/debate

  ❑ *Deliberative Democracy*

- Computational interlocutors mining web for arguments [1] and engaging human interlocutors e.g. in educational technologies for enhancing student learning and reasoning skills

  ❑ E.g., *E-Clinic* application plays role of consultant on ward rounds challenging student diagnosis/treatment plan

  ❑ E.g. Socratic search/argumentation engine engaging politics/philosophy students

[1] *Computers that can argue will be satnav for the moral maze.* New Scientist, September 2016

Applications of Argumentation-based Dialogue for Scaffolding Human Reasoning

- ❑ Filtering algorithms = technological amplifications of evolutionary dispositions to seek supporting arguments/evidence and ignore arguments against

- ❑ Computational interlocutors exposing users to opposing views/arguments to help dismantle echo chambers and burst filter bubbles ?

- ❑ But in these contexts people not motivated to consider opposing views/evidence

- ❑ Need **early** educational interventions to inculcate more interactive/dialectical engagement with information - success may depend on extent to which dispositions are "hard wired" by evolution or cognitive gadgets "installed by nurture"

(see Celia Hayes. *The Cultural Evolution of Thinking* &
Catarina D. Novaes. *The enduring enigma of reason*". Mind & Language 2018.)

Dialogical support for aligning AI and Human Values

- As AI becomes more powerful and autonomous they are likely to achieve goals in ways misaligned with human values (and hence potentially harmful *)

- *Future of Humanity Institute (Oxford)*, *Centre for the Study of Existential Risk, Open AI, MIT ….* all working on value loading/alignment problem

- State of the art = *cooperative inverse reinforcement learning* – AI learns reward function of human through dialogical interaction – will require dialogical models of distributed reasoning

- Facilitating joint human and AI reasoning could enable **better moral decision making**, e.g. leveraging superior epistemic and causal reasoning of AI and reasoning about preferences and values of human

* *Superintelligence: Paths, Dangers, Strategies.* Nick Bostrom (head of *Future of Humanity Institute*, Oxford University)

Challenges for Dialectical formalisations of Non-monotonic Reasoning

- ❑ ASPIC+ = **general** framework for argumentative formalisations of non-monotonic reasoning *

- ❑ ASPIC+ does not satisfy all criteria for rationality and rationality postulates that are satisfied assume agents with unbounded resources

- ❑ ASPIC+ does not accommodate typical dialectical uses of arguments

- ❑ Our work is focused on D-ASPIC+ - a dialectical framework that **is fully rational under resource bounds** (currently under review – IJCAI 2020)

- ❑ Currently we have a dialectical account of a special instance of ASPIC + - *classical logic argumentation* - that is fully rational under resource bounds

* S. Modgil, H. Prakken. *A General Account of Argumentation and Preferences*. In: Artificial Intelligence (AIJ) 2013