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

$|\sim_{NmL}\alpha|$

$NmL = Preferred Subtheories, Prioritised Default Logic, Defeasible Logic, Logic Programming$
Argumentative Formalisations of Non-monotonic Reasoning

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

$\models_{\text{NmL}} \alpha$

$\text{NmL} = \text{Preferred Subtheories}$

$\text{Prioritised Default Logic,}$

$\text{Defeasible Logic, Logic Programming} \ldots$

**Args**
e.g., $A = \{a, a \rightarrow b\}, b$

**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 Information

Argument Framework
< Args, Defeats >

$\alpha \models_{NmL} \neg \alpha$

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

iff

$\alpha$ is the claim of a winning (justified) argument

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

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

Dialectical status of locution (claim $\alpha$) is winning iff $\Delta \vdash \neg \alpha$

Dialogical support for:
1) Enhancing quality and scope of human reasoning;
2) Enabling joint human and AI reasoning


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

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 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

---