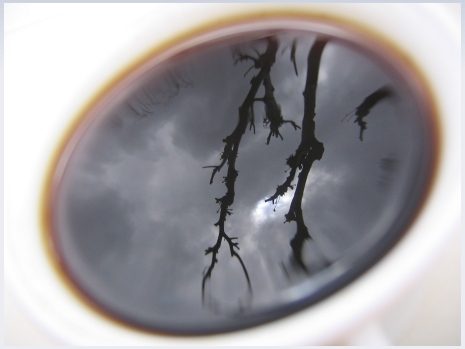
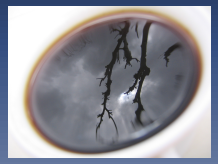


From Argument Games to Persuasion Dialogues

Nicolas Maudet (aka Nicholas of Paris)

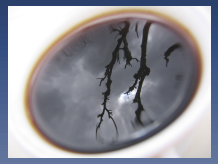
08/02/10 (DGHRCM workshop)
LAMSADE

Université Paris-Dauphine



Main sources of inspiration for this talk

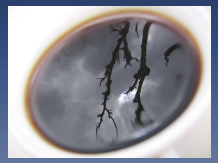
- "Process and Policy: Resource-Bounded non Demonstrative Reasoning" (R. Loui, 1998)
- "Formal Systems for Persuasion Dialogue" (H. Prakken, 2006)



Objective of the talk

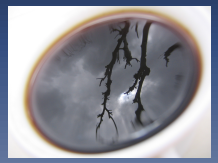
- provide an overview of the use of "dialogue games" from an AI perspective;
- briefly present recent models and issues discussed in the community.

I start by giving some general reasons as to why these models proved popular in (different branches of) AI...



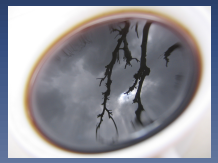
From the AI and Law perspective

- logic alone does not allow to model the burden of proof switch;
- DG are useful because they emphasize the procedural aspect of reasoning;
- descriptive/prescriptive view.



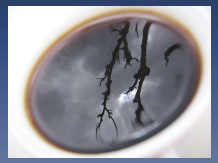
From the dialogue systems perspective

- dialogue models purely based on intentional notions have theoretical and practical (complexity of planning) limits;
- DG are useful because they provide stereotypic patterns of dialogue (as identified in conversation analysis);
- descriptive/prescriptive view.



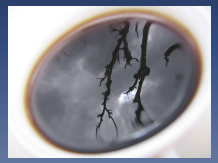
From the software agents perspective

- semantics of communication languages used by software agents cannot be checked if they refer to (private) notions, eg. mental states;
- DG are useful because they focus on the notion of commitments, a publicly verifiable notion.
- prescriptive view.



Overview of the rest of the talk

1. Preliminaries: abstract argumentation systems
2. Formal models of persuasion
3. Argumentation games
4. Agents playing with persuasion games



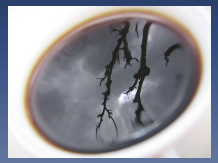
Abstract Argumentation Systems

Most of the work discussed in this talk is connected to the **abstract argument systems** as initiated by Dung (1989).

- Very popular framework allowing to capture different types of non-monotonic reasoning.
- Abstracts away from the actual content of arguments.

Definition (argumentation system)

An argumentation system is a pair $\langle A, R \rangle$ where A is a (finite) set of arguments, and R an attack relation between arguments ($R \subseteq A \times A$)



Abstract Argumentation Systems

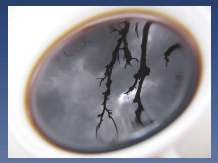
Given an argumentation system, we would like to specify what are the coherent points of view, ie. what it means for sets of arguments to be acceptable.
Acceptability is typically based on two requirements:

- **internal stability**: the set of arguments does not contain arguments that attack each others;
- **external stability**: the set of arguments should ... (many possible definitions).

A central notion here is that of collective defense:

Definition (Collective defense)

A set of arguments S collectively defends a, for any $b \in A$ st. bRa, there exists $c \in S$ st. cRb

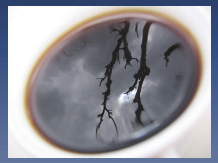


Different semantics

Many different semantics have been proposed in the literature, using different notions of external stability, in particular:

- **stable**: S must attack every arguments outside S;
- **admissible sets**: S must defend all its arguments
- **prefererred**: $\subseteq$-maximal among admissible sets
- **grounded**: least fixed point of the function returning the set of arguments defended by a given set of arguments.

Note that some of these semantics allow for different extensions, in which case we classically distinguish **credulous** vs. **skeptical** acceptability of arguments.



Fleshing out the abstract system

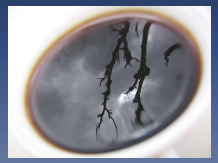
Of course at some point you need to specify what constitutes your basic argument entities. One classical way to do so is to assume that they are built from a (monotonic) underlying logic. E.g.:

- arguments may be pairs $\langle H, h \rangle$, where H is the support and h the conclusion, st. (i) H entails h and (ii) H is minimal.

As for attack relations following (Pollock,1986) we may have:

- **undercutters** (when the conclusion of an argument contradicts one of the premises of the support of another argument), or
- **rebuttals** (when conclusions directly contradict).

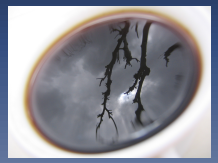
Note that non-monotonicity comes from the interaction between arguments, not from the underlying logic.



Audiences and preferences

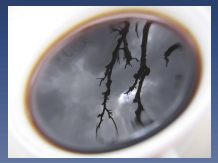
The need to incorporate preferential information exogenously given on the arguments quickly emerged (giving rise to many approaches, eg. value-based argumentation framework (Bench-Capon et. al, 2002):

- Suppose there are **labels** attached to the different arguments, eg. tagging which topic/value/criteria/source they refer to.
- Suppose there is a preference relation over these labels, eg. a linear order of the different criteria
- Now the attack relation can be refined so that it takes this preferential information into account (eg. discard attacks of less preferred arguments).
- An **audience** defines one possible ordering over these values.
- Arguments may now be **objectively** acceptable (regardless of the audience considered), or **subjectively** acceptable.



Overview of the rest of the talk

1. Preliminaries: abstract argumentation systems
2. Formal models of persuasion
3. Argumentation games
4. Agents playing with persuasion games

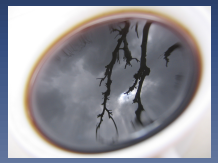


Basic elements

(Prakken, 2006) gives the following basic elements required to specify dialogue games:

- a **topic language** $\mathcal{L}_t$;
- a **communication language** $\mathcal{L}_c$;
- a **dialogue purpose**;
- a set of **participants** and a set of **roles**. Each participant is potentially equipped with a (possibly inconsistent) **knowledge base** ($\Sigma_i \subseteq 2^{\mathcal{L}_t}$), and is assigned a **commitment store** (CS_i).
- a **context**, the shared knowledge presupposed and invariant during the dialogue.

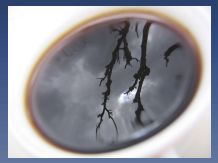
Note that no commitments and knowledge bases need not coincide.



Dialogue rules

Dialogue rules regulate the dialogue in several ways. The following types of dialogue rules can be distinguished:

- a set of **effect rules**: defines how the commitment stores are affected by the moves
- a **protocol** which assign to legal finite dialogues the allowed moves to play next.
  Termination occurs when no move is allowed.
  Turn-taking specifically specifies who should speak next.

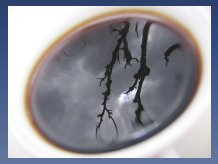


Persuasion dialogues

In the specific context of persuasion dialogues, we can instantiate some of these parameters:

- we distinguish some propositions (**topics**) of $\mathcal{L}_t$ which will be debated upon;
- each participant may be, wrt to each topic, either a **proponent**, an **opponent**, or a neutral **third-party**.
- outcome rules specify who wins on each topic.

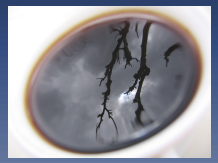
In general, a winner (resp. loser) on a topic must (resp. not) have the discussed proposition in his *CS* at the end of the dialogue.
In **pure persuasion** (two-player) dialogues, the protocol terminates as soon as the proponent sees (all) his topic(s) conceded by the opponent (or vice-versa).



Specific models

The proposed model is general enough to cater for a wide range of dialogue systems as proposed in the literature, in particular in the works of C. Hamblin (1971), J. McKenzie (1980), or Walton and Krabbe (1995), and more recently by many researchers in AI. Now all the models make different choices regarding different issues, we briefly discuss:

- what degree of flexibility is allowed?
- what are the arguments exchanged?
- how is consistency checked?

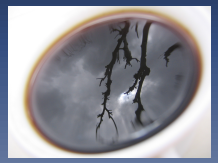


What flexibility is allowed?

The first level of flexibility is defined at the level of **turn-taking**.

- does the protocol strictly alternate between players?
- does it allow for several moves to be made in a row?
- does the turn switch as a result of some more sophisticated function (eg. depending on the "status" of the current dialogue?)

Two others aspects are involved: are there typical dialectical obligations attached to moves? how is relevance of the moves constrained?

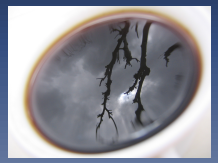


Locutions and their typical replies

Prakken (2006) summarizes the typical replies that persuasion systems allow as **replies** to a given move:

Locutions	Replies
claim ψ	why ψ, claim $\overline{\psi}$, concede ψ
why ψ	ψ since S (or: claim S), retract ψ
concede ψ	
retract ψ	
ψ since S	why ϕ ($\phi \in S$), concede ϕ ($\phi \in S$)

If the underlying logic permits, you may also have counter-arguments consisting of concessions followed by attacks on some of the premisses.

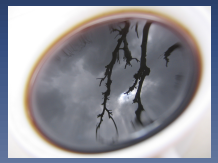


Relevance

By requiring each move to be an immediate reply to the one uttered just before, we do not allow **backtracking** replies.
Prakken (2005) attaches to each move advanced in the dialogue a **dialogical status** (in/out), in particular to the main claim.

- Replies are partitioned as surrenders/attackers.
- A move is *in* if it is surrendered or if all its attackers are *out*.
- A move is relevant iff it switches the dialogical status of the initial claim.

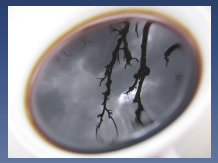
Note that this still does not allow for "cross-examinations", where a sequence of moves is required to establish relevance.



How is consistency checked?

Typically, consistency checking of the commitments made so far can be either:

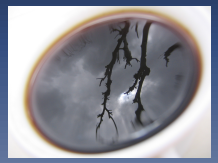
- enforced at the level of dialogue rules.
  Eg. it is not permitted to claim something if you are already commited on the opposite proposition.
- left to the agents themselves
  Eg. the DC system includes a *resolve* move requiring the partner to solve an inconsistency in his CS.
  Eg. WK's system specifies that if the CS of an agent (x) implies a claim made by the other agent, which x is not committed to, then x must either concede the claim of retract of the implying commitments.



What are the exchanged arguments?

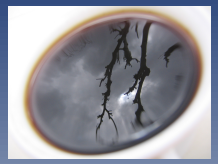
Finally, there is the question of what is the content of arguments being exchanged between agents. In particular we may or not:

- allow for partial arguments (enthymemes)
  Eg. McKenzie's DC system allows incomplete arguments to be given, but commit the agent to the missing (material implication) premise.
- allow step-by-step revelation of arguments instead of fully constructed arguments.



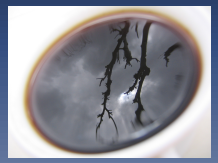
Overview of the rest of the talk

1. Preliminaries: abstract argumentation systems
2. Formal models of persuasion
3. Argumentation games
4. Agents playing with persuasion games



Dialectical proof-theories for argumentation semantics

Argument games have been proposed as dialectical proof-theories allowing to determine the acceptability of a given argument.
These games constitute very restricted and rigid form of dialogues.
Different games have been proposed for the different semantics of Dung, either for credulous or skeptical acceptability.
In the next slide, I give the example of an argument game for the grounded semantics.

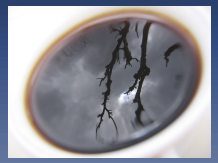


Argument game for grounded semantics

The following argument game has been proposed as a proof-theory for grounded semantics acceptability (Prakken, Sartor, 1997).

- the proponent first puts forward the argument she wants to prove;
- then, respecting strict turn-taking, at step i:
  - the opponent replies with an argument which attacks $arg(m_{i-1})$
  - the proponent replies with an argument which attacks $arg(m_{i-1})$ and such that $arg(m_{i-1})$ does not attack $arg(m_i)$ (not repeating previous moves).

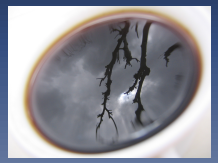
The argument is acceptable iff proponent has a winning strategy to defend this argument.



Properties of argumentation games

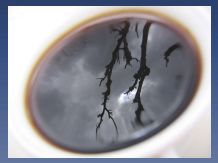
Of course we require the dialectical procedure to be sound and complete wrt. the semantics being considered.
But other properties may be pursued:

- Characterization of semantics allowing a **certain class of games** to be used (eg. coherent systems allow two-party immediate response protocols).
- We may be interested in the **length** of the obtained dialogues. As such dialogues can serve as certificates to a given solution of a argumentation decision problem, we expect dialogues to be generally long (= requiring exponentially many steps) for those problems lying above NP (see Dunne, 2001, 2003).
- Robustness against **dynamification** of the protocol.



Overview of the rest of the talk

1. Preliminaries: abstract argumentation systems
2. Formal models of persuasion
3. Argumentation games
4. Agents playing with persuasion games

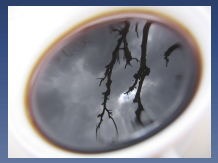


Assumptions

Study the properties of some games assuming they are played by certain types of agents (Amgoud et. al, 2000-present). Now we need to detail the relation between knowledge bases and commitments.

- Agents are equipped with **knowledge bases** Σ_i which remain static during the dialogue;
- The underlying reasoning machinery uses Dung's argumentation systems and grounded semantics;
- Commitment stores are subsets of knowledge bases;
- To construct their arguments, each agent can make use of his Σ and of the CS of the other agent.

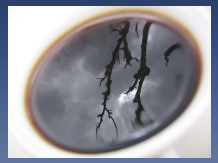
(Below $\mathcal{A}(.)$ denotes the set of arguments which can be constructed from a given set of propositions, and $\mathcal{S}(.)$ for those that are acceptable.



Attitudes, rationality rules

Each agent adopts an **attitude** wrt. assertion and acceptance, which specifies what to "rationally" assert or accept given what can be inferred from his knowledge base.
For instance:

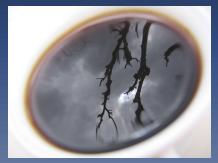
- A **credulous** agent accepts a proposition p as long as it's backed by an argument.
- A **cautious** agent accepts a proposition p if he is unable to construct an acceptable argument for $\neg p$
- A **skeptical** agent accepts a proposition p if he has an acceptable argument for it.



An example mechanism

The following mechanism defines both the protocol and the strategy of the agents:

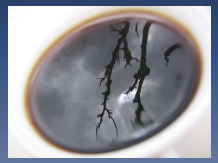
1. A assert p;
2. if acceptance attitude allows: B accepts p;
   else if assertion attitude allows B asserts $\neg p$;
   else B challenge p
3. if B asserts $\neg p$ goto 2 (with roles exch. and $\neg p$ insted of p;
4. if B has challenged:
   (i) A asserts S (support of p)
   (ii) for each $s \in S$, goto (2) in turn.



Defining the outcomes of the game

- **knowledge outcomes** for P:
  $O_k(P|C) = \{p|\exists a \in \mathcal{A}(\Sigma_P \cup CS_C) \text{ st. } p = conc(a)\}$
- **joint knowledge outcomes**:
  $O_k(P \wedge C) = \{p|\exists a \in \mathcal{A}(\Sigma_P \cup \Sigma_C) \text{ st. } p = conc(a)\}$
- **committed outcomes** for P:
  $O_c(P|C) = \{p|\exists a \in \mathcal{A}(CS_P) \text{ st. } p = conc(a)\}$
- **joint committed outcomes**:
  $O_c(P \wedge C) = \{p|\exists a \in \mathcal{A}(CS_F \cup CS_C) \text{ st. } p = conc(a)\}$

... and similarly for **acceptable** outcomes (denoted $O_k^a(P|C)$, etc.)

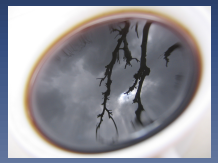


Some interesting properties

- An agent can win with a proposition not in $O_k^a(P \wedge C)$.

Example:

$\Sigma_P = \{\neg c, a, a \rightarrow b\}$
$\Sigma_C = \{\neg c \rightarrow \neg b\}$
P asserts b, which C accepts, although $b \notin \mathcal{S}(\Sigma_P \cup \Sigma_C)$



Some interesting properties

- An agent can win with a proposition not in $O_k^a(P \wedge C)$.

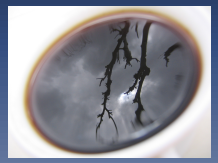
Example:

$\Sigma_P = \{\neg c, a, a \rightarrow b\}$
$\Sigma_C = \{\neg c \rightarrow \neg b\}$
P asserts b, which C accepts, although $b \notin \mathcal{S}(\Sigma_P \cup \Sigma_C)$

- The order of locutions may make a difference on the outcome.

Example:

$\Sigma_P = \{a, a \rightarrow b, b \rightarrow c, a \rightarrow f, f \rightarrow c\}$
$\Sigma_C = \{b, b \rightarrow \neg(f \rightarrow c), f \rightarrow \neg(b \rightarrow c)\}$
If P asserts $\langle\{a, a \rightarrow b, b \rightarrow c\}, c\rangle$, and C accepts c.
If P asserts $\langle\{a, a \rightarrow f, f \rightarrow c\}, c\rangle$, but then C can contradict both arguments that P might build for c.



...

- The extent to which the outcome of the dialogue is predetermined to be in $O_k^a(P \wedge C)$ is an interesting measure;
- The following properties can be defined for protocols: **reachability** or **convergence** to certain outcomes.
- Minimal requirement (Prakken, 2006): reachability of joint knowledge outcomes (note that the mechanism proposed above does not satisfy this).
- But what do we want exactly?
  - **fairness** both agents might have the possibility to win;
  - **control** the outcome should reflect the beliefs of the agents (as given in their bases)