Normative Multi Agent Systems
"Sanction based obligations in a qualitative decision theory"

Guido Boella
Università di Torino
Leendert van der Torre
Vrije Universiteit

Obligations in MAS

- Obligations play an important role in the "programming" of multi agent systems.

  They stabilize the behavior of a multiagent system, and thus play the same role as intentions do for single agent systems …

Explicit representation of norms or implicit ?

"An obligation holds when there is an agent A, the *normative* agent, who has a goal that another (or more than one) agent B, the *bearer* agent, satisfy a goal G and who, in case he knows that the agent B has not adopted the goal G, can decide to perform an action Act which (negatively) affects some aspect of the world which (presumably) interests B. Both agents know these facts"

[Boella and Lesmo, 2002]

Violations…

- The agent cannot do anything for the norm.
- The plans to achieve it achieves a low utility.
- A plan not fulfilling the obligation but inducing the *normative* agent to believe otherwise.
- A plan not fulfilling the obligation but which makes the sanction impossible to be applied
- The *bearer* bribes (or menaces) him
- …

Carmo and Jones 2002

- *Normative systems* are “sets of agents (human or artificial) whose interactions can fruitfully be regarded as norm-governed; the norms prescribe how the agents ideally should and should not behave [...]. Importantly, the norms allow for the possibility that actual behaviour may at times deviate from the ideal, i.e. that violations of obligations, or of agents rights, may occur”

Normative “agents”

- We attribute mental states to normative systems such as legal or moral systems, a proposal which may be seen as an instance of Dennett’s *intentional stance* [Dennett, 1987]:
- Agent-style characteristics: autonomy, proactivity, social awareness and reactivity - mental attitudes: such as beliefs, desires and intentions

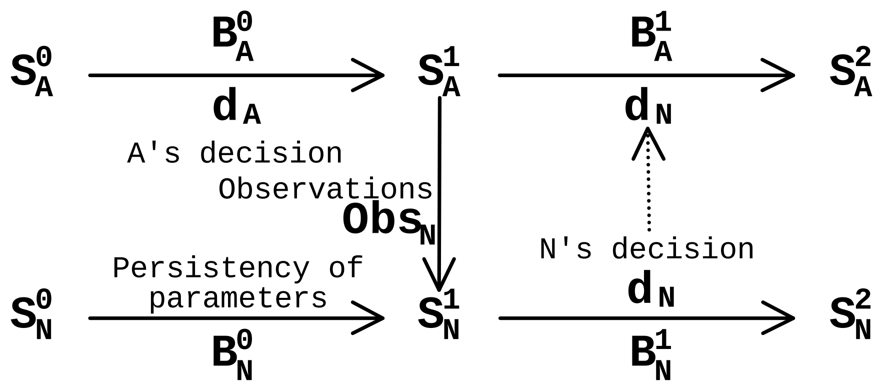
Social order

- [Castelfranchi, 2001] *multiagent* systems as "*dynamic social orders*": patterns of interactions among interfering agents "*such that it allows the satisfaction of the interests of some agent*".
- "a shared goal, a value that is good for everybody or for most of the members"
- Social order requires *social control*, "*an incessant local (micro) activity* of its units, able to restore or reproduce the regularities prescribed by norms"

Obligations

1) The content of the obligation is a desire and goal of N and N wants that A adopts this goal.
2) N has the desire and the goal that, if the obligation is not respected by A, a prosecution process is started to recognized if the situation 'counts as' a violation and that, if a violation is recognized, A is sanctioned.
3) Both A and N do not desire the sanction: for A the sanction is an incentive to respect the obligation, while N has no immediate advantage from sanctioning.

Recursive modeling



Decisions

- Let *A={a1,a2, ...}*, *N={n1,n2, ...}* and *P={p1,p2, ...}* be three disjunct sets of propositional variables, i.e. $A \cap N = \emptyset$, $A \cap P = \emptyset$, and $N \cap P = \emptyset$. A literal is a variable or its negation.
- A *decision set* is a tuple $\langle dA, dN \rangle$ where *dA* is a set of literals of *A* (the decision of agent A) and *dN* is a set of literals of *N* (the decision of agent N).

Epistemic states

- Let P^0, P^1 and P^2 be the sets of propositional variables defined by $P^i=\{p^i \mid p\in P\}$.
- LA, LAP^1, ... the propositional languages built up from A, $A\cup P^1$, ...
- The *epistemic state* is a tuple

  $\langle s^A{}_0, s^A{}_1, s^A{}_2, s^N{}_0, s^N{}_1, s^N{}_2\rangle$ where $s^A{}_0$ and $s^N{}_0$ are sets of literals of LP^0, $s^A{}_1$ and $s^N{}_1$ are sets of literals of LAP_1), and $s^A{}_2$ and $s^N{}_2$ of LNP^2

Rules

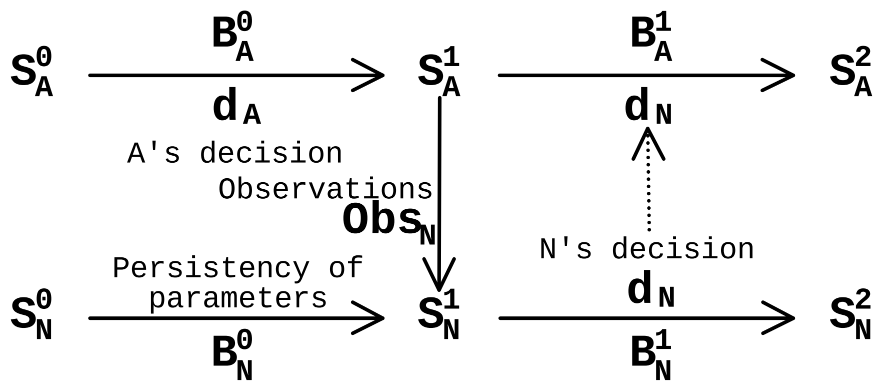
- Two sets of *belief rules* are used to calculate the expected consequences of decisions and two sets of *desire and goal rules* are used to evaluate the consequences of decisions.
- A rule is an ordered pair of sentences
- $l_1\wedge...\wedge l_n \rightarrow l$, where $l_1,...,l_n,l$ are literals of this language.

Mental state

- The *mental state* is a tuple $\langle B^A_1, B^A_2, B^N_1, B^N_2, D^A, G^A, D^N, G^N \rangle$
- B^A_1 and B^N_1 are sets of rules of LAP^0P^1,
- B^A_2 and B^N_2 are sets of rules of $LANP^0P^1P^2$,
- D^A, G^A, D^N and G^N are sets of rules of $LANP^0P^1P^2$.

- The set of observable propositions *Obs* is a subset of $A \cup P^1$. The *expected observation* of N in state s^A_1 is
  $Obs_N = \{p \mid p \in Obs \text{ and } p \in s^A 1\} \cup$
  $\{\neg p \mid p \in Obs \text{ and } \neg p \in s^A_1\}$.

Recursive modeling



Observations

- The set of observable propositions *Obs* is a subset of $A \cup P^1$. The *expected observation* of N in state $s^A{}_1$ is
  $Obs_N = \{p \mid p \in Obs$ and $p \in s^A{}_1\} \cup$
  $\{\neg p \mid p \in Obs$ and $\neg p \in s^A{}_1\}$.

Consequences

- For rational agents, the epistemic state is a consequence of applying belief rules to the previous state, together with persistence of the previous state

Respecting mental states

For s a state, f a set of literals of $LANP^1$ and R a set of rules, let $max(s,f,R)$ be the set of states:

1. $\{\{l1,...,ln\} \cup f \mid l_{i,1} \wedge ... \wedge l_{i,mi} \rightarrow l_i \in R$ for $i=1...n$ and $l_{i,j} \in s \cup f$ for $j = 1...m_i$ and $\{l_1,...,l_n\} \cup f$ consistent $\}$
2. $S' = \{s \in S \mid \exists s' \in S$ such that $s \subseteq s'\}$
3. $max(s,f,R) = \{s' \cup s'' \mid s' \in S'$ and $s'' = \{l^i \in s \mid l^i \in P^i$ and $\neg l^{i+1} \notin s'\}\}$

Respecting

$\langle s^A{}_0, s^A{}_1, s^A{}_2, s^N{}_0, s^N{}_1, s^N{}_2 \rangle$ *respects* $\langle dA, dN \rangle$, Obs_N and $\langle B^A{}_1, B^A{}_2, B^N{}_1, B^N{}_2, D^A, G^A, D^N, G^N \rangle$

if

$s^A{}_1 \in max(s^A{}_0, dA, B^A{}_1)$,

$s^A 2 \in max(s^A{}_0 \cup s^A{}_1, dN, B^A{}_2)$,

$s^N{}_1 \in max(s^N{}_0, Obs_N, B^N{}_1)$

$s^N{}_2 \in max(s^N{}_0 \cup s^N{}_1, dN, B^N{}_2)$.

Unfulfilled mental states

$U(R,s)=\{l_1 \wedge \ldots \wedge l_n \rightarrow l \in R \mid$
$\{l_1, \ldots, l_n\} \subseteq s$ and $l \notin s\}$

The *unfulfilled mental state description* of A is the tuple $\langle U_A{}^{DA}, U_A{}^{GA}, U_A{}^{DN}, U_A{}^{GN}, U_N \rangle$ where $U_A{}^{DA}=U(D_A,s^A)$, $U_A{}^{GA}=U(G_A,s^A)$, $U_A{}^{DN}=U(D_N,s^A)$, $U_A{}^{GN}=U(G_N,s^A)$, and $U_N = \langle U_N{}^{DN}, U_N{}^{GN} \rangle$ is the unfulfilled mental state of N: $U_N{}^{DN}=U(D_N,s^N)$, $U_N{}^{GN}=U(G_N,s^N)$.

Agent characteristics

$\langle \geq^A{}_B, \geq_A, \geq^N{}_B, \geq^N \rangle$ where $\geq^A{}_B$ is a transitive and reflexive relation on the powerset of B^A, $\geq_A$ is a transitive and reflexive relation on the powerset of $D^A \cup G^A \cup D^N \cup G^N$, $\geq^N$ is a transitive and reflexive relation on the powerset of B^N, and $\geq^N{}_B$ is a transitive and reflexive relation on the powerset of $D^N \cup G^N$.

Respecting mental states and beliefs

- For *s* a state, *f* a set of literals in $LANP^1$, *R* a set of rules, and a transitive and reflexive relation on *R* containing at least the superset relation, let $max(s,f,R, \geq)$ …

Agent types (from BOID)

1. if AT = *stable* then $U^A{}_N \geq U'^A{}_N$ iff $U^{GA}{}_A \geq U'^{GA}{}_A$ and if $U^{GA}{}_A \geq U'^{GA}{}_A$ and $U'^{GA}{}_A \geq U^{GA}{}_A$ then $U^{DA}{}_A \geq U'^{DA}{}_A$
2. if AT = *unstable* then $U^A{}_N \geq U'^A{}_N$ iff $U^{DA}{}_A \geq U'^{DA}{}_A$ and if $U^{DA}{}_A \geq U'^{DA}{}_A$ and $U'^{DA}{}_A \geq U^{DA}{}_A$ then $U^{GA}{}_A \geq U'^{GA}{}_A$
3. if AT = *OGNonly* then $U^A{}_N \geq U'^A{}_N$ iff $Obl(U^{GN}{}_A) \geq Obl(U'^{GN}{}_A)$ where $Obl(U^{GN}{}_A)$ is the set of obligations of A
   (the rules $l1 \wedge ... \wedge ln \rightarrow l \in G^N$ such that $l \in A$).

Optimal decisions

$\langle dA,dN \rangle$ *minimal for* N if for every other decision set $\langle dA,dN' \rangle$ with unfulfilled mental state $U'N = UN$ then $dN \subseteq dN'$.

$\langle dA,dN \rangle$ is *optimal for* N if it is minimal for N and for every expected state description $s'N$ of a N minimal decision set $\langle dA,dN' \rangle$ there is an expected state description sN of $\langle dA,dN \rangle$ such that $s^N \geq s'^N$.

A decision specification is *conflict free* if the optimal decision set for A is unique

Anderson's reduction of modal logic

- $O(p)=NEC(\neg p \rightarrow V)$:
  if *p* is obliged, then it is necessarily the case that the negation of *p* implies the violation constant *V*.
- However many violations are not sanctioned.
- He later interpreted it as 'something bad has happened'.
- We read it as 'the absence of *p* counts as a violation' (as in Searle's construction of social reality)

Obligations: O(A,N,a)

Agent A believes to be *obliged* to decide to do *a* ($a \in$ A an ought-to-do obligation) iff A believes that:

1. $\rightarrow a \in D^N \cap G^N$: Agent N desires and has as a goal that *a* and wants A to adopt *a* as a goal.
2. $\exists v \in N\ \neg a \rightarrow v \in D^N \cap G^N$: If $\neg a$ then N has the goal and the desire to recognize it as a violation *v*.
3. $\rightarrow \neg v \in D^N$: N desires that there are no violations.
4. $\rightarrow \neg v >^N \neg a \rightarrow v$

Obligations with sanction O(A,N,a,s)

Agent A believes to be *obliged* to decide to do *a* with sanction *s* (a decision variable in *N*) iff:

1. Agent A believes to be obliged to decide to do *a*, as defined above.
2. $v \rightarrow s \in D^N \cap G^N$: A believes that if *v* then agent N desires and has as a goal that it sanctions A.
3. $\rightarrow \neg s \in D^N$: agent A believes that agent N desires not to sanction $\neg s$.
4. 4. $\rightarrow \neg s \in D^A$: Agent A has the desire for $\neg s$, which expresses that it does not like to be sanctioned.

Sanction as parameters

Agent A believes to be *obliged* to decide to do *a* with sanction *s* (a parameter in P^2 to be achieved by agent N) iff:

1. Agent A is obliged to decide to do *a* with sanction *s* as defined above, but now with *s* a parameter in P^2.
2. $\exists n \in N \quad n \rightarrow s \in B^N$: agent A believes that agent N has a way to apply the sanction.

Example: O(A,N,a,s)

$s^A_0 = \emptyset,\ B^A = \emptyset,\ G^A = \emptyset,$

$D^A = \{\rightarrow\neg a, \rightarrow\neg s\},$

$\geq^A = \{\rightarrow\neg a\} > \{\rightarrow\neg s\}$

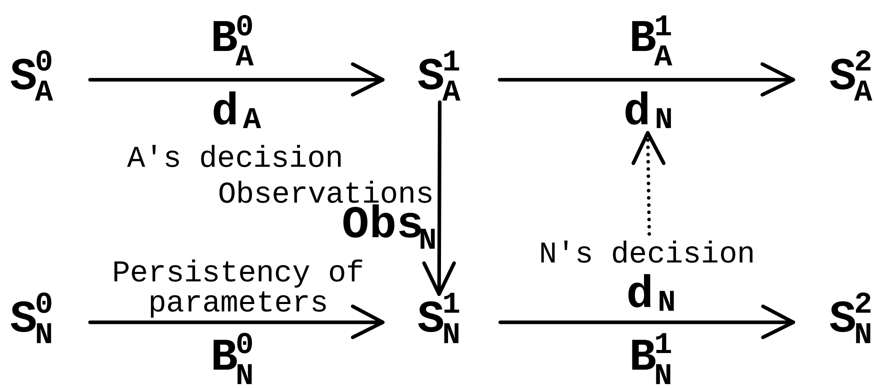
$s^N_0 = \emptyset,\ Obs_N = A \cup P^1,\ B^N = \emptyset,$

$G^N = \{\rightarrow a,\ \neg a \rightarrow v,\ v \rightarrow s\},$

$D^N = \{\rightarrow a,\ \neg a \rightarrow v,\ v \rightarrow s,\ \rightarrow\neg v,\ \rightarrow\neg s\},$

$\geq^N = \{\rightarrow\neg v\} > \{\neg a \rightarrow v\}, \{\rightarrow\neg s\} > \{v \rightarrow s\}$

Recursive modeling



decision set: $<dA = \{\neg a\}, dN = \emptyset>$

$s^A{}_1=\{\neg a\}, s^N{}_1 = \{\neg a\}, s^A{}_2 = \emptyset, s^N{}_2 = \emptyset$

Unfulfilled mental states

$\mathrm{U}^A= \emptyset$

$U^N=\{\neg a \rightarrow v\}$

decision set: $<dA = \{\neg a\}, dN = \{v,s\}>$

$s^A{}_1=\{\neg a\}, s^N{}_1 = \{\neg a\}, s^A{}_2 = \{v,s\}, s^N{}_2 = \{v,s\}$

Unfulfilled mental states

$\mathrm{U}^A=\{\rightarrow\neg s\}$

$U^N= \{\rightarrow\neg v, \rightarrow\neg s\}$

decision set: $<dA = \{a\}, dN = \varnothing>$

$s^A{}_1 = \{a\}, s^N{}_1 = \{a\}, s^A{}_2 = \varnothing, s^N{}_2 = \varnothing$

Unfulfilled mental states

$\mathrm{U}^A=\{\rightarrow\neg a\}$

$U^N= \varnothing$