Computer Models of Social Practices
The Constitutive Interpretation

Richard Evans

PT-AI 2013, Oxford

1 The Constitutive View of Social Practices

1.1 The Regulative View of Social Practice

Research in multi-agent systems typically[1] assumes a *regulative* model of social practice. In this model, the agent starts with a given set of goals, and a given set of actions for achieving these goals. Social practices are then introduced afterwards as a way of achieving coordination when there are multiple agents whose activity can come into conflict. In this regulative model, a social practice provides a *restriction* on the set of (antecedently given) actions so as to satisfy the (antecedently given) set of goals. For example, in a world containing cars but no driving regulations, agents are free to drive on either side of the road. To prevent collisions, we introduce driving regulations, insisting that everyone drives on the left hand side of the road. We accept this limitation on our freedom because it helps us satisfy our goal of survival.

The alternative *constitutive* view of social practices denies that actions or goals can be specified in advance, prior to and independently of the practices the agent is participating in. The constitutive view claims that:

- There are some actions that are only available because one is participating in a social practice with a certain structure
- There are certain goals that are only available within the social practice that institutes the value

1.2 Some Actions are only Available in Certain Practices

Examples:

- You can swing a peculiarly shaped piece of wood without participating in any particular practice - but this action will only constitute a *strike* if you are participating in a game of baseball
- I can raise my hand whenever I like, but this only counts as *voting for the motion* within the institution of voting
- I am free to say "I do"at any moment, but these sounds only constitute *getting married* within a wedding ceremony

[1] See [5] or [13]. For an explicit acknowledgement of the importance of constitutive conditionals in multi-agent simulations, and the beginning of an analysis, see [3].

Clarification 1: there are some actions I cannot perform alone because of some contingent fact about myself. So, for example, I cannot move the sofa on my own. But perhaps, if I was stronger or more manly, I could. The constitutive claim much stronger than that: the claim is that there are certain actions which necessarily cannot be done unless they are achieved within a particular practice.

Clarification 2: the new action that the practice enables is not a new type of physical action. It isn't like the case of Spiderman, who was bitten by a radioactive spider and suddenly had new physical capacities (e.g. the ability to climb walls). Rather, the new actions that are made available by practices are *re-interpretations* of previously-achievable actions. Using Searle's counts-as formulation [10]:

x counts as y in context c

The social practice provides a context c in which the already-achievable action x is now also the performance of y. So, to take a very well-worn example, moving the piece of wood from one square to another *counts as* moving the knight to king's-bishop 3 in the context of the game of chess.

1.3 Some Goals are only Available in Certain Practices

It isn't just actions that are constituted in practices. Properties of objects, more generally, are constituted through practice. So, for example:

- This wooden object counts as a *knight* in this game of chess
- This ring of stones counts as a *boundary* for this tribe [10]

Some of our wants are desires that objects have such practice-instituted properties. So, for example, you can only want to protect your king-side if you are playing chess. The property that you want to satisfy (a well-defended king-side) is instituted in the practice, and does not make sense without it.

It is important to have a healthy array of examples. Games, with their explicit rules, carefully worded to minimize ambiguity, are only one rather special type of practice. Here is a different sort of example: the goal of being an honourable man is only available within a certain (old-fashioned and largely defunct) set of social practices which institute the notion of honour. Or, to take a less familiar case: one can only want to offer a *pokala* within the complex social practices of the Kula ring.

1.4 Computer Models of Constitutive Practice

This talk will describe computer models of constitutive social practices satisfying both these points: there are both actions and goals which are *only available* because of the practices in which they are instituted.

One way to see the need for a constitutive view of practice is to consider the vast array of actions I could possibly do. Sitting here right now, I could lend

the stranger on my left 10 pounds; I could tell the other stranger on my right that Paris is the capital of France; I could ring up my wife and enumerate the prime numbers... With an infinite number of actions available to me, why am I not overwhelmed with choice? How do I ever find the time to make a decision?

In the constitutive view of practice, every action is embedded in a practice: the agent does not see the action as available unless he is already participating in the practice that makes it visible. The agent isn't overwhelmed by an infinite number of choices because *he only sees the actions that are provided by the social practices he is in.*

1.5 Social Practices as Foundational

The constitutive view, as outlined so far, is based on the picture of an agent who starts off already able to represent, reason and plan, and with a set of physical capacities, and the practices provide him with a way of expanding his capacities and goals.

But this work is based on a stronger, more controversial understanding of constitutive social practice: the agent's ability to represent, reason and plan is itself constituted by the practices he is participating in. Intentionality itself is one of the (clusters of) capacities that social practices enable.

Consider Searle's distinction in [9] between *intrinsic* and *derived* intentionality. According to him, people have intrinsic intentionality, while pieces of paper (and computers) have merely derived intentionality: the fact that the writing on the piece of paper means something is only true because there are agents who (intrinsically) mean something by that same sentence when they produce/interpret it. The more controversial understanding of constitutive practice claims that the intentionality of *people* is also derivative. The only thing that has intrinsic intentionality is the *practice.*

More specifically, when an agent means something by a sentence, that is only because he is participating in a practice in which that sentence means what it does. This is an example of a *mediating relation.*

Mediating Relations Some relations can be best understood by splitting them in two, and inserting an intermediate entity between the relata:

In these cases, the explanation is of the form:

$$\forall x, y\ A(x, y) \text{ because } \exists z\ B(x, z) \wedge C(z, y)$$

For example:

- x is the aunt of y because there is a person z who is the sibling of x and the parent of y
- x is a logical (proof-theoretic) consequence of y iff there is a proof z such that z starts with y and ends with x
- x is married to y because there has been a wedding ceremony z in which x was groom in z and y was bride in z

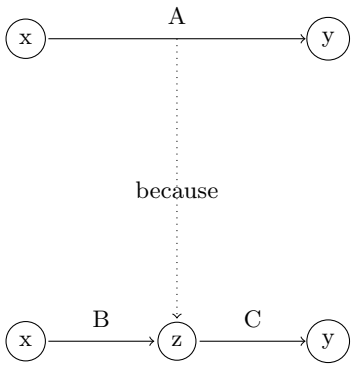

Fig. 1: Mediating Explanations

One of the central claims in [2] is that intentional states should be explained in terms of participation in practices:

> Expressions come to mean what they mean by being used as they are in practice, and intentional states and attributes have the contents they do in virtue of the role they play in the behavioral economy of those to whom they are attributed.

This is another example of a mediating explanation. Having an intentional state is a relation between an agent and an intentional state which is best understood via a mediating entity - a social practice:

– x has intentional state y because there is a practice z such that x participates in z and z institutes y

1.6 Summary

So far, I have outlined three levels of increasing commitment to social practices:

1. Social practices as *restrictions* on (antecedently given) actions to satisfy (antecedently given) goals
2. Social practices as *amplificative*: ways of providing extra actions and goals to agents who are already intentional (representing, reasoning, planning)
3. Social practices as *foundational*: providing the structure needed for all intentional activity

To place some names on these positions:

1. Much work in multi-agent AI research (e.g. Moses and Tennenholtz [5], Shoham [13]) work from the first assumption
2. Searle[10] and Rawls[6] work from the second assumption
3. Brandom[2] explicitly argues for the third position

The computer models described start from the third, strongest, most controversial assumption. If we want to evaluate the claim that activity in general is made possible through social practice, then we should start with the hardest case first. The hardest case is intentionality. If we can make it plausible that intentionality itself can be instituted through social practice, then the other cases will be relatively straightforward.

2 Modelling the GOGAR

2.1 Introduction

The GOGAR (**G**ame **O**f **G**iving and **A**sking for **R**easons) is a social practice which enables participants to make assertions by producing sounds. If a parrot says "Nice weather we're having", she has not *asserted* that the weather is nice - she has merely produced some noises. But if an agent utters these sounds as a participant in the GOGAR, it counts as *asserting* that the weather is nice. GOGAR is the practice which turns sounds into assertions.

The GOGAR keeps track of who has said what (the commitments). It divides the commitments into two groups: those that are entitled, and those that are not.

Claims are entitled by default, but a claim can lose its entitlement if it is challenged. A claim can be challenged by a non-propositional speech-act (the incredulous raising of an eyebrow), or by a propositional speech-act: by asserting a proposition which is incompatible with it.

When a claim is challenged, it loses its entitlement. But it can get its entitlement reinstated if it is justified by other assertions. These justifications are themselves just other assertions, which can themselves be challenged by other incompatible assertions - in which case, the justifications will themselves need further justifications to reinstate their entitlement - and these further justifications can themselves be challenged, and so on.

The rest of this section will describe computer models of the GOGAR. The description of the GOGAR in [2] is remarkably precise, making it relatively straightforward to implement what he described. But we will see that we will have to depart from Brandom at one crucial point.

2.2 Basic Definitions

Given a background set $\mathcal{X}$ of agents and $\mathcal{S}$ of sentences, then a debate-state $\mathcal{D}_x$ according to a particular agent $x \in X$ is a tuple $(\mathcal{A}, \mathcal{C}, \mathcal{E}, \mathcal{I})$, consisting of:

1. A set $\mathcal{A} \subseteq \mathcal{X} \times \mathcal{S}$ of assertions
2. A set $\mathcal{C} \subseteq \mathcal{S} \times \mathcal{P}(\mathcal{S})$ of commitment-preserving inferences in horn-clause form
3. A set $\mathcal{E} \subseteq \mathcal{S} \times \mathcal{P}(\mathcal{S})$ of entitlement-preserving inferences in horn-clause form
4. A set $\mathcal{I} \subseteq \mathcal{P}(\mathcal{P}(\mathcal{S})$ of sets of incompatibility sets

Different agents will have different interpretations of the same debate. The set $\mathcal{A}$ of assertions will vary between different agents: one might not have heard an assertion if he misheard, was out of earshot, or was not paying attention. Different agents may have different understandings of the inferential relations: one agent may think that p is a commitment-preserving consequence of q, while another may not. There can be similar disagreements about entitlement-preserving relations and incompatibility relations: one agent may think that, for example, free will and determinism are incompatible propositions, while another may think they are compatible.

At the heart of the GOGAR is a function which computes, given a particular understanding of the debate $\mathcal{D}_x = (\mathcal{A}, \mathcal{C}, \mathcal{E}, \mathcal{I})$ according to a particular agent x, the subset of assertions which x thinks are entitled:

$$\text{Entitled}_x(\mathcal{A}, \mathcal{C}, \mathcal{E}, \mathcal{I}) \subseteq \mathcal{A}$$

All claims are entitled by default. A claim loses its entitlement if it is challenged. The entitled claims are claims that either remain unchallenged - or have been challenged, but have also been successfully justified. To say that a claim is not entitled is to make a claim with normative consequences: the claims that are not entitled are the ones that the asserter *should* justify (or retract).

2.3 Incompatibility, Commitment and Entitlement

In *Making It Explicit* (p.194), Brandom says

> Two claims are incompatible if commitment to one precludes entitlement to the other

But incompatibility is a relation between sentences, while commitment is a relation between a speaker and a sentence. If p and q range over sentences, and x and y range over agents, there are two possible interpretations:

1. $\mathcal{I}(p, q)$ iff $\mathcal{A}(x, p)$ precludes Entitled(y, q)
2. $\mathcal{I}(p, q)$ iff $\mathcal{A}(x, p)$ precludes Entitled(x, q)

In the first interpretation, one speaker's assertion can affect the entitlement of another speaker's assertion. In the second interpretation, one speaker is making both assertions.

The first interpretation is too strong. It means that if x asserts p, I can prevent x ever being entitled to p just by asserting an incompatible proposition q. This interpretation is clearly is not what Brandom means.

The second interpretation is the one that John MacFarlane ascribes to Brandom. But it is, I submit, too weak. Suppose x asserts p and y asserts q, where p and q are incompatible. According to this second interpretation, both x's claim that p and y's claim that q can *both* retain their entitlement, even though they are incompatible. This means that the assertions of one speaker can have no effect whatsoever on the entitlement of another speaker's assertions! They are

closed off from affecting each other. But one person's claim can only count as a *challenge* to another person's claim if the former can affect the entitlement of the latter.

So, in one way, our second interpretation is much too weak. But in another way it is too strong. Suppose x asserts p and q (where again p and q are incompatible). Suppose further that x provides some very compelling arguments for p, but none for q. Now he is committed to q, so if commitment to q precludes entitlement to p, then he can never get entitlement reinstated for p, *no matter how compelling* the justifications he provides for p. The only way, according to this interpretation, that he can get entitlement for p is if he *retracts* q.

The second interpretation is the one that MacFarlane implemented in his own implementation of GOGAR [4]. But his implementation has exactly the issue outlined above, that different speakers' claims cannot affect each others' entitlements:

```
Welcome to the game of giving and asking for reasons,
a simulation of the linguistic scorekeeping dynamics
described in chapter 3 of Robert Brandom's book
Making It Explicit (Harvard University Press, 1994).

(c) 2006 John MacFarlane

For a list of sample commands, type help

GOGAR> Bob asserts A is red

GOGAR> Ann asserts A is blue

Bob's score on Ann
Commitments:  {A is colored, A is blue}
Entitlements: {A is colored, A is blue}
Incompatibles:

Bob's score on Bob
Commitments:  {A is red, A is colored}
Entitlements: {A is red, A is colored}
Incompatibles:

GOGAR> Bob asserts A is blue
Bob's score on Ann
Commitments:  {A is colored, A is blue}
Entitlements: {A is colored, A is blue}
Incompatibles:

Bob's score on Bob
```

```
Commitments:  {A is red, A is colored, A is blue}
Entitlements: {}
Incompatibles: {A is red, A is blue}
```

First Bob asserts "A is red". Then Ann asserts "A is blue"[2]. At this point, despite the fact that they have challenged each other with directly incompatible assertions, neither of the two claims has lost entitlement. Finally, Bob asserts "A is blue". It is only when a speaker contradicts *himself* that he loses entitlement. This simple example shows the problems with this second interpretations: direct challenges go entirely unnoticed.

As neither of the above interpretations are satisfactory, we are compelled to look elsewhere. The simplest alternative is:

3. $\mathcal{I}(p, q)$ iff Entitled(x, p) precludes Entitled(y, q)

Two claims are incompatible if entitlement to one precludes entitlement to the other (no matter who said them).

The difference between this interpretation and Brandom's is that he sees entitlement as a collective version of *justification*: just as two people can believe incompatible claims, and both be justified in their beliefs, just so two people can be entitled to their incompatible assertions. In the proposed alternative interpretation, by contrast, entitlement is a collective version of *knowledge*. Imagine a social practice in which the monadic predicate "It is known that p" is prior to the dyadic "x knows that p". In this alternative interpretation, entitlement to p represents "It is known that p". Since two incompatible propositions cannot both be known, two incompatible propositions cannot both be entitled.

In the rest of this section, I describe a computer implementation of GOGAR which assumes this alternative interpretation of the relation between incompatibility and entitlement.

2.4 Computing Entitlement

To ease exposition, we make two simplifying assumptions:

1. We ignore who said what, treating an assertion as a simple sentence (ignoring who uttered it)
2. We ignore the commitment-preserving inferences, and just assume the set of assertions is closed under the commitment-preserving relation

So now a debate $\mathcal{D}_x$ is just a triple $(\mathcal{A}, \mathcal{E}, \mathcal{I})$, consisting of:

1. A set $\mathcal{A} \subseteq \mathcal{S}$ of assertions
2. A set $\mathcal{E} \subseteq \mathcal{S} \times \mathcal{P}(\mathcal{S})$ of entitlement-preserving inferences in horn-clause form
3. A set $\mathcal{I} \subseteq \mathcal{P}(\mathcal{P}(\mathcal{S})$ of sets of incompatibility sets

[2] The implicit assumption is that A is monochromatic.

Example 1. For example

$$\mathcal{A} = \{p, \neg p, q\}$$
$$\mathcal{E} = \{p \leftarrow \{q\}\}$$
$$\mathcal{I} = \{\{p, \neg p\}\}$$

Figure 1 shows the state of the debate in Example 1. Arrows indicate entitlement-

Fig. 2: Example 1

preserving rules and dotted boxes represent incompatibility sets. Entitled claims are drawn in circles.

Example 2.

$$\mathcal{A} = \{p, \neg p, q, r\}$$
$$\mathcal{E} = \{p \leftarrow \{q\}, \neg p \leftarrow \{r\}\}$$
$$\mathcal{I} = \{\{p, \neg p\}\}$$

Fig. 3: Example 2

Example 3.

$$\mathcal{A} = \{p, \neg p, q, r, s\}$$
$$\mathcal{E} = \{p \leftarrow \{q\}, p \leftarrow \{s\}, \neg p \leftarrow \{r\}\}$$
$$\mathcal{I} = \{\{p, \neg p\}\}$$

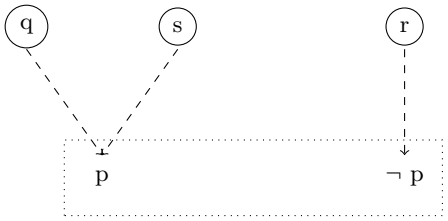

Fig. 4: Example 3

Example 4.

$$\mathcal{A} = \{p, \neg p, q, r, s, t\}$$
$$\mathcal{E} = \{p \leftarrow \{q\}, p \leftarrow \{s\}, \neg p \leftarrow \{r\}, r \leftarrow \{t\}\}$$
$$\mathcal{I} = \{\{p, \neg p\}\}$$

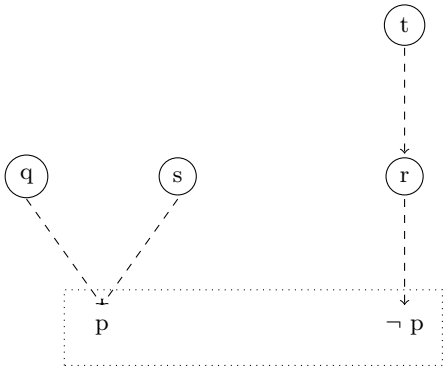

Fig. 5: Example 4

Computing the entitled claims involves, at the procedural level:

1. Start by assuming all the claims are entitled
2. Remove entitlement from all incompatible subsets
3. Compute the immediate consequences of the currently entitled claims
4. Remove entitlement from all incompatible subsets
5. Repeat steps 3 and 4 until no more propositions are added to the set of entitled claims

Redescribing this procedure at the functional level, let

$$\text{Inc}(s) = \{x \in s \mid \exists y_1, ..., y_n \in s \text{ such that } \{x, y_1, ..., y_n\} \in \mathcal{I}\}$$

Let

$$\phi(s) = s - \text{Inc}(s)$$

Define $Cn_1(s)$ as the set of immediate consequences of s according to the horn-clauses in $\mathcal{E}$:

$$Cn_1(s) = s \cup \{p \in s \mid (p \leftarrow q) \in \mathcal{E} \wedge q \subseteq s\}$$

Now define a function $N : \mathcal{P}(T) \rightarrow \mathcal{P}(T)$:

$$N = \phi \cdot Cn_1$$

Note that ϕ and N are not monotonic. Now define a sequence of entitlement sets $E_0, E_1, ...$ where:

$$E_0 = \phi(\mathcal{A})$$
$$E_{n+1} = N(E_n)$$

Now, although N is not monotonic, we have $E_n \subseteq E_{n+1}$ for all n. So, if $\mathcal{S}$ and $\mathcal{E}$ are finite, this sequence converges, and we define

$$\text{Entitled}(\mathcal{A}, \mathcal{E}, \mathcal{I}) = \bigcup_{i \geq 0} E_i$$

Example 5. Given:

$$\mathcal{A} = \{p, \neg p, q, r, s, t, u\}$$
$$\mathcal{E} = \{p \leftarrow \{q, r\}, \neg p \leftarrow \{s\}\}\}$$
$$\mathcal{I} = \{\{p, \neg p\}, \{t, u, s\}\}$$

The computation of Entitled involves:

$$E_0 = \{q, r\}$$
$$E_1 = \{q, r, p\}$$
$$E_2 = \{q, r, p\}$$
$$...$$
$$\text{Entitled}(\mathcal{A}, \mathcal{E}, \mathcal{I}) = \{q, r, p\}$$

Example 6. A slightly more complex example, consider:

$$\mathcal{A} = \{p, \neg p, q, \neg q, r, \neg r, s\}$$
$$\mathcal{E} = \{p \leftarrow \{q\}, q \leftarrow \{r\}, r \leftarrow \{s\}\}\}$$
$$\mathcal{I} = \{\{p, \neg p\}, \{q, \neg q\}, \{r, \neg r\}\}$$

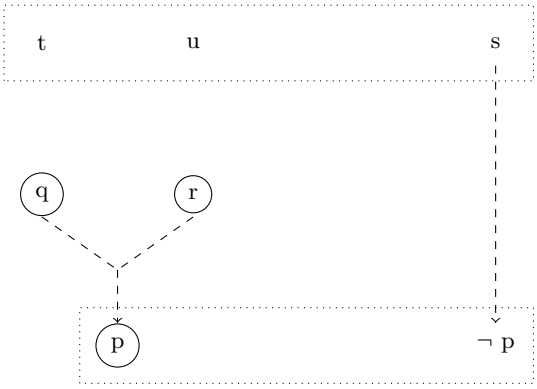

Fig. 6: Example 5

The computation of Entitled involves:

$$\begin{aligned} E_0 &= \{s\} \\ E_1 &= \{s, r\} \\ E_2 &= \{s, r, q\} \\ E_3 &= \{s, r, q, p\} \\ E_4 &= \{s, r, q, p\} \\ &... \\ \text{Entitled}(\mathcal{A}, \mathcal{E}, \mathcal{I}) &= \{s, r, q, \neg p\} \end{aligned}$$

Entitlement-Preserving Inferences Do Not Always Preserve Entitlement in the Presence of Conflicting Claims One noteworthy aspect of this way of computing entitlement is that it allows the possibility that the following are simultaneously true:

- There is an entitlement-preserving inference from q to p: $p \leftarrow \{q\} \in \mathcal{E}$
- q is entitled: $q \in \text{Entitled}(\mathcal{A}, \mathcal{E}, \mathcal{I})$
- p is not entitled: $p \notin \text{Entitled}(\mathcal{A}, \mathcal{E}, \mathcal{I})$

For example, consider the following case:

Example 7.

$$\begin{aligned} \mathcal{A} &= \{p, \neg p, q, r\} \\ \mathcal{E} &= \{p \leftarrow \{q\}, \neg p \leftarrow \{r\}\} \\ \mathcal{I} &= \{\{p, \neg p\}\} \end{aligned}$$

Here, the entitled claims are:

$$\text{Entitled}(\mathcal{A}, \mathcal{E}, \mathcal{I}) = \{q, r\}$$

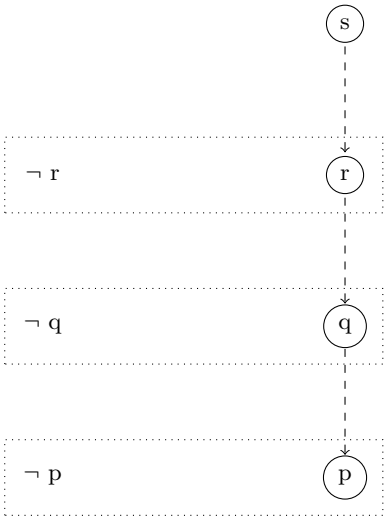

Fig. 7: Example 6

Fig. 8: Example 3

Now, it might seem odd (or worse, plain wrong) to allow an entitlement-preserving inference that does not automatically transfer entitlement from premise to conclusion (and contrapositively, transfer lack of entitlement from conclusion to premise). But this is to misunderstand the nature of an entitlement-preserving inference: it does preserve entitlement by default (see Cn_1) - but only *in the absence of countervailing factors*. This is a non-monotonic logic: other commitments may preclude entitlement from the conclusion even if the premise is entitled.

An alternative approach would be to apply modus-tollens to entitlement and percolate non-entitlement backwards from conclusions to premises. So if there is a rule $p \leftarrow \{q\} \in \mathcal{E}$ and p lacks entitlement, then q should lose entitlement too. But what does this alternative approach do when there are multiple conjuncts in the body of the clause? When we have, say, $p \leftarrow \{q, r\}$, and p loses entitlement, which of q and r lose entitlement? Or do they both lose entitlement? Isn't that too harsh? Well, does the *conjunction* of p and q lose entitlement? How does that

affect the entitlement of the individual atoms? Surely if $p \wedge q$ lacks entitlement, then p and q cannot both be entitled? Well, then, which has lost entitlement?

We avoid these murky questions if we reject modus-tollens on entitlement and keep with the original idea: entitlement-preserving inferences preserve entitlement in the *absence* of other commitments - but when there are other commitments in play, Cn_1 is only a default entitlement preserving link which may be overridden by the incompatibility-removing ϕ operator.

Entitlement as an Abstract Normative Status Entitlement, as defined here, is an abstract property which is used to determine which statements require justification. The way entitlement is defined abstracts away from many of the historical details of how the claims were uttered:

- Making the same statement *more than once* has no effect whatsoever on the set of entitled claims
- The *order* in which claims are made has no effect whatsoever on the set of entitled claims
- *Who* is making which claims has no effect whatsoever on the set of entitled claims. If you claim p and $\neg p$, then entitlement-wise this is no different from if you claim p and I claim $\neg p$: in both cases, both claims lack entitlement

Computing the Extra Assertions Needed to Make a Claim Entitled Suppose, in a debate, a certain claim lacks entitlement and we want to reinstate its entitlement. Typically there will be various different sets of claims one could add which would restore its entitlement. These sets can be calculated as follows. Given a debate $\mathcal{D} = (\mathcal{A}, \mathcal{E}, \mathcal{I})$ involving a total set $\mathcal{S}$ of propositions, the possible justifications $\mathcal{J}$ of a claim p are defined as:

$$\mathcal{J}(p) = \{c' \in \mathcal{P}(\mathcal{S} - \mathcal{A}) \mid p \in \text{Entitled}(c' \cup \mathcal{A}, \mathcal{E}, \mathcal{I}) \wedge \neg\, \exists c'' \in \mathcal{P}(\mathcal{S} - \mathcal{A})\ c'' \subset c'\}$$

In other words, the possible justifications of p are the minimal subsets c' of $\mathcal{S}$ that can be added to $\mathcal{A}$ such that p is entitled in $(c' \cup \mathcal{A}, \mathcal{E}, \mathcal{I})$.

Example 8. Suppose we have

$$\begin{aligned}\mathcal{A} &= \{p, \neg p, q, \neg q, r, \neg r, s, t, u\}\\ \mathcal{E} &= \{p \leftarrow \{q, r\}, \neg p \leftarrow \{s\}, \neg p \leftarrow \{t, u\}\}\\ \mathcal{I} &= \{\{p, \neg p\}, \{q, \neg q\}, \{r, \neg r\}\}\end{aligned}$$

Then the various justifications of $\neg p$ are:

$$\mathcal{J}(\neg p) = \{\{\neg q, s\}, \{\neg r, s\}, \{\neg q, t, u\}, \{\neg r, t, u\}\}$$

It is a consequence of the definition of entitlement that a debate can never get into a hopeless position. Any claim in any debate is *redeemable* in the sense that for all propositions p that are claimed in debate $(\mathcal{A}, \mathcal{E}, \mathcal{I})$ involving propositions $\mathcal{P}$, there are extensions $(\mathcal{A}' \supseteq \mathcal{A}, \mathcal{E}' \supseteq \mathcal{E}, \mathcal{I}' \supseteq \mathcal{I})$ in $\mathcal{P}' \supseteq \mathcal{P}$ such that p is entitled in $(\mathcal{A}', \mathcal{E}', \mathcal{I}')$.

3 Modelling Pragmatic Factors

3.1 The Burden of Proof

One norm lies at the heart of the GOGAR:

> If x is committed to p, and p lacks entitlement, then x should either justify or retract p.

In situations where two agents have made incompatible assertions, neither of which is justified, they *both* have to justify (or retract) their claims. This model contrasts with the traditional concept of the "burden of proof". According to this concept, there is at most one person who has the "burden of proof" at any moment. It may be sometimes difficult to assess who has it, and different people may disagree, but nevertheless there is at most one person who can have it.

The model proposed above is more equanimous. There are often cases (see Examples 2, 3 and 4) where multiple agents have to justify their claims. There is no single agent who the burden of proof falls on. In these cases, any attempt to single out an individual as particularly responsible is an attempt to impose a *power relation* on an essentially symmetrical situation.

3.2 Norms and Power Relations

At the heart of the GOGAR lies the norm:

> If x is committed to p, and p lacks entitlement, then x should either justify or retract p.

But norms do not work by magic. They need to be articulated, monitored and enforced by the activities of individual agents.

When there is a norm that x should justify p, who should articulate, monitor and enforce that norm? There are two broad ways this can happen:

- Another y can insist that x justify his claim by playing *high-status* to x. (For example, y says to x: "You really need to justify your claim that p").
- x himself can enforce the norm by playing low-status to another y. (For example, x says to y: "Oh dear, my claim that p has been challenged. I really need to justify it").

The terms *high status* and *low status* come from Keith Johnson's *Impro*. These are power relations in a short-term local situation[3]. In Johnson's sense, a pauper can play high-status to a king while he is bossing him around. (Humour often arises from unexpected status games).

The high/low status-game is implemented as a separate social practice, running concurrently with the GOGAR, which monitors the state of the GOGAR and provides appropriate status-related affordances. The status-game can be in one of three states:

[3] As opposed to social status (e.g. upper class) which are long-term and (at least in some societies) difficult to change within one's own life-time

- Neutral: no status game is currently being played
- High/Low(x,y): agent y is playing high to agent x
- Conflict(x, y): agent x and y are in conflict as a result of trying to play high-status to each other

There are different affordances available in the different states. In the Neutral state, whenever there is a claimer x who has claimed something which is not entitled, there is an opportunity for

- x to play low to anyone
- another y to play high to x

In the High/Low(x,y) state, the affordances available include:

- y remind x that x needs to justify his/her claim
- x look worried
- x get annoyed with y (if x does not like being low status)

So, for example, if x and y have made incompatible assertions p and q, and neither p nor q has been justified, then they can both play high-status on each other. Suppose y plays high status to x. Now if x is also a habitual high-status player, he will not enjoy being made to play low-status, and will get annoyed.

3.3 Turn-Taking and Conversational Salience

In an earlier implementation of GOGAR, the agents had a good understanding of when their claims had been challenged, and what sort of claims to make to justify their assertions - but they had no understanding of the conversational context in which their claims were embedded. If an agent had one of his claims challenged earlier, but then the conversation had moved on, the agent would go back, relentlessly, to his previously challenged assertion. He had no understanding of whose turn it was to speak, or the current focus (or foci) of the conversation.

To address this, our latest implementation combines a model of giving-and-asking-for-reasons with a model of conversational salience (based on Sacks, Schegloff and Jefferson's seminal paper on conversational turn-taking[8]). Conversational turn-taking is itself modelled as a social practice, containing norms describing who should speak next, on what topic, and with various mechanisms for repairing conversational blunders. Now, the agents understand both what has been said and *when it is appropriate to respond.*

The turn-taking practice involves two core concepts:

- the selected speaker (if any)
- the selected topic (or topics)

The selected speaker is the agent (if any) who should speak next. Certain types of utterance directly determine a selected speaker. For example:

- x says to y: "Can you please stop standing on my foot?"

Others indirectly determine a selected speaker. If x has previously asserted p and y makes a claim which is incompatible with q, then this is a challenge to x's previous claim and x is the selected speaker. But sometimes, there is no selected next speaker and anybody can speak next. For example:

– x (addressing the group in general): "Has anybody seen my hat?"

Each utterance is about one or more topics. The most recent utterance determines the selected topics.

The turn-taking practice enforces the following rules:

– If there is a selected speaker, then he should speak next. Other people speaking out of turn constitutes an interruption.
– If there is no selected speaker, anyone may speak next.
– The next utterance should involve one of the selected topics. If the selected speaker cannot continue any of the selected topics, he must preface his utterance with a preamble connecting it to a previous topic, or preface it with a conventional way of clearing selected topics ("anyway..."). Failure to respect this rule constitutes an interruption.

4 Evaluation, Limitations and Further Work

4.1 An AI Architecture Inspired by Philosophical Insight

Austin once wrote [1]:

> In the history of human inquiry, philosophy has the place of the central sun, seminal and tumultuous: from time to time it throws off some portion of itself to take station as a science, a planet, cool and well regulated, progressing steadily towards a distant final state.

He believed that one day[4], the heated debates surrounding philosophy of language would cool down, and the hard-won accumulated insights would start to build into a science of language[5].

The computer model described here started by taking seriously the idea that philosophical insight can serve as direct inspiration for AI architectures. A number of the fundamental architectural features were based on (controversial) philosophical insights. I have already dwelt at length with three such claims:

– certain actions are constituted in certain practices
– certain goals are constituted in certain practices
– intentionality, in particular, is constituted in a particular practice (GOGAR)

[4] He believed it would happen during the 21st century

[5] When he said a science of language, he did not just mean a science of formal linguistics (a la Chomsky and formal grammar), but a science of language *use*.

I shall briefly describe one other philosophical claim that fundamentally informed the architecture. Typically multi-agent simulations of social practices model the world as a collection of objects[6]. This simulation, by contrast, takes the Tractarian idea seriously that the world is *everything that is the case*: the entire simulation state is defined as a set of sentences in a formal language. Choosing a declarative representation of simulation state has been shown to have certain significant practical advantages when building a complex simulation: it is significantly easier to visualise, debug, and serialise the simulation state[7]. Time and again, hard-won philosophical insights informed the fundamental architectural decisions.

4.2 Limitations and Further Work

So far, I have built simulations of simple debates using the architecture described above. But these are only toy examples and initial explorations. There is a huge amount more to do. I will focus on two aspects in particular:

- supporting language entry and exit moves
- agents making their inferential relations explicit

4.3 Supporting Language Entry and Exit Moves

Sellars [12] coined the term "language entry move" for an inference from a perception to an assertion. He used the term "language exit move" for a transition from an assertion to an action. The computer model of GOGAR implemented so far has neither language-entry nor language-exit moves.

- Language-entry moves. The agents start off with a given set of beliefs. This set does not expand or contract during simulation.
- Language-exit moves. The computer agents do not *act* on their beliefs (apart from asserting or justifying them).

A richer simulation would model the way agents acquire information, how a debate can change someone's mind, and how their beliefs can affect their subsequent actions (as Marx famously insisted on).

[6] In an object-oriented representation, each object is a cluster of facts, related to other objects via pointers.

[7] The formal language used to represent the world was not traditional predicate logic, but a modal language designed to model social practices. This modal logic, Exclusion Logic, is itself inspired by the Sellars/Brandom thesis that material incompatibility is prior to logical negation. In this logic, you can express directly the fact that "x is blue" and "x is red" are incompatible.

4.4 Making Inferential Relations Explicit

In the GOGAR, agents typically have different understandings of the inferential relations. They will have different understandings about which claims are incompatible, commitment-preserving, and entitlement-preserving. Their different understandings of the inferential relations leads to different understandings of the entitlements. For example, suppose x believes free-will and determinism are incompatible, while y believes they are compatible. Consider the following exchange:

1. x: 'We have free-will'
2. y: 'All events are entirely determined'
3. x: 'Yes, that is also true'

By the third claim, at the point at which x agrees with y, x (the compatibilist) believes both 1 and 2 are entitled. y (who is an incompatibilist) thinks both claims have lost entitlement.

In the current implementation, such differences cannot be resolved because they cannot be made explicit. But consider one possible continuation:

4. y: 'Huh? You have a contradicted yourself'
5. x: 'No I haven't'
6. y: 'Yes you have. You claimed that we have free well, and that all events are entirely determined. These two claims are incompatible.'
7. y: 'No - they are not incompatible.'
8. x: 'Yes they are. It is part of the meaning of an event being freely willed that you could have done otherwise - but if determinism is true, you could not have done otherwise.'

In the continuation, the debate is raised to the meta-level. Instead of disagreeing about first-order issues (free-will versus determinism), they are now disagreeing about whether x has contradicted himself, and whether or not two claims are incompatible. There is an incompatibility at the meta-level between the claims "x has contradicted himself" and "x has not contradicted himself". There is also an incompatibility between "p and q are compatible" and "p and q are incompatible". By the 7th claim, y has challenged the incompatibility of free-will and determinism. This incompatibility claim (which was previously implicit in x) is now challenged.

Recall the first-level norm:

> If x is committed to p, and p lacks entitlement, then x should either justify or retract p.

There is a related norm at the meta-level:

> Once an inferential relation has been made explicit and lost entitlement, it should no longer be used in inference until its entitlement is restored.

So x is stymied until he can provide justification for the incompatibility. It is only when he makes the 8th claim that entitlement is restored.

Meta-level debating involves making the various inferential relations explicit. We have seen an example where an incompatibility relation was made explicit. Here is an example where an entitlement-preserving relation is made explicit and challenged:

1. x: p
2. y: $\neg p$
3. x: q
4. y: "You need to provide a justification for p."
5. x: "No I don't - p is already justified by q."
6. y: "On the contrary - q may be true but it does not support p."

Here, they disagree about whether p is entitled, and then proceed to disagree about whether q is an entitlement-preserving reason for p.

When we add this sort of meta-level debating to the GOGAR, agents will be able to change their mind - not only about what they believe - but also about what they *mean*.

References

1. Austin, J. L.: Ifs and Cans. Proceedings of the British Academy 1956. Reprinted in Collected Papers 1979.
2. Brandom, R.: Making It Explicit. Harvard University Press (1998)
3. Jones, A., Sergot, M.: A Formal Characterisation of Institutionalised Power. Journal of the IGPL (1996)
4. MacFarlane, J,: GOGAR. http://johnmacfarlane.net/gogar.html
5. Moses, Y., Tenenholtz, M.: On Computational Aspects of Artificial Social Systems. Proc 11th DAI Workshop (1992)
6. Rawls, J.: Two Concepts of Rules. The Philosophical Review, Vol.64 (1955)
7. Sacks, H.: Lectures on Conversation. Kluwer (1989)
8. Sacks, H., Schegloff, E., Jefferson, G.: A Simplest Systematics for the Organization of Turn-Taking for Conversation. Language, Vol. 50 (1974)
9. Searle, J, R.: The Rediscovery of the Mind. MIT Press (1992)
10. Searle, J, R. Making the Social World: The Structure of Human Civilization. Oxford ; New York: Oxford University Press (2010)
11. Sellars, W.: Empiricism and the Philosophy of Mind. Harvard University Press (1997)
12. Sellars, W.: Some Reflections on Language Games. Philosophy of Science, Vol 21 (1954)
13. Shoham, Y.: Multiagent Systems. Cambridge University Press (2008)
14. Wittgenstein, L.: Philosophical Remarks. University of Chicago Press. 1975.
15. Wittgenstein, L.:Tractatus Logico-Philosophicus. Routledge and Kegan Paul, 1961.