Introduction

Disclaiming Games

Trust and Restraint

Application: Biscuits

References

Day 5: Cooperativity, Reputation and Hedging

Eric McCready

Aoyama Gakuin University

August 14, 2015

Earlier (Day1/2): external mechanisms can enforce cooperativity.

- Jury!

But also reputation (as discussed).

- Question: How to protect one's reputation???

Optimally: be cooperative, speak the truth.

- Problem: fallibility.
- Solution: assertion/talk in such a way as to reduce culpability.

One possibility: hedge.

A simple example: advertising.

- One common kind of advertising involves pictorial images.
- However, it is often the case that the actual object lacks some, or even most, characteristics of the object in the image.

Puzzles:

- Why do we let images influence us when we know they are not accurate?
- Since this is common knowledge, why do sellers give pictorial advertisements at all?
  - Limit to representational case.

One relevant factor:

- Disclaimers.

(1) shashin-wa imeeji desu
photo-Top image Cop
'The picture is an 'image'.' (Japanese)

(2) Actual product may differ (from picture)
Actual contents may vary.

But these phrases also deepen the puzzle.

- Why trust after an explicit statement of lack of responsibility?
- And given this, why use such a statement at all?

Is this just a capitalist phenomenon? No. *Hedging*!

(3) a. *I suspect that* it is cold outside.
b. *I might be wrong, but* Palin is not going to be elected.
c. *This might not be true, but* she doesn't really care about you.

- These expressions 'guard' the speaker from being accused of saying falsehoods. (esp. *shields* Prince *et al.* 1982: (3b) and (3c)).
- Claim: the use of these expressions is similar to that of disclaimers in advertising.
- Goal in hedging is to avoid being held responsible for utterance content if false, goal of a disclaimer to avoid being held responsible if the actual product is unsatisfactory.

Aims of today's lecture:

- What is the use of expressions like these?
- Formal account of function of 'shield'-type hedges
- Extensions: other kinds of hedging, other kinds of content

Plan:

- Brief discussion of cooperation in evolutionary theory
- Modeling (shield) hedging
- Hedging nonentailed content

Cooperation in evolution

A longstanding problem in evolutionary theory (e.g. Nowak 2006).

- Goal in interaction (assumed): maximization of fitness
- Competition for resources leads to expectation of noncooperation
  - Prisoner's Dilemma situations: unilateral noncooperation leads to higher fitness for individual
  - but universal noncooperation leads to lowered fitness for all
- But: parents-offspring, eusocial insects, pack behavior, . . .
- Cooperation quite widespread

How to explain these phenomena?

Some possibilities:

1. Group selection
   - Evolution acts at non-individual level
2. Kin selection
   - Individuals act to maximize fitness of individuals with shared genes
3. Repeated games
   - In the long term, cooperation leads to higher utility

(1,2) seem not immediately relevant to linguistic behavior. (3)!

Repeated games

Iterated PD: cooperation the best move (when number of iterations not known).

- Repeating games gives incentives to cooperate.
- If repeating game gives a higher payoff for cooperation eventually ...
- Over multiple iterations, cooperative behavior 'evolves.'

Buying watches:

- If multiple interactions expected, cooperative move is best.
- Seller loses any benefit for initial L move soon if buyer plays $\neg b$ thereafter.

Note the same thing happens without advertising ...

- Advertising works to start the cooperative cycle, if believed.

Finally, noncooperative games of communication.

- Speaker payoffs depend on whether communication is successful, i.e. whether he has convinced the hearer to believe him and on whether *H* picks a particular state.
- In repeated games, punishment by *H* is available: use a purely probabilistic strategy picking most likely state.
  - In purely cooperative settings, more severe punishment available.
- This strategy will not yield much payoff for the speaker in the long run, only whatever the probability of the optimal state is times speaker's payoff.
- With enough repetitions, again, cooperation is best for the speaker in general

Work on evolution of cooperation shows the utility of this type of strategy (Nowak, 2006).
Three possible strategies in iterated PD:

- Grim: cooperate until defected on
- Tit-for-tat: copy opponent's previous move
- ⋆ Generous TFT: copy opponent's previous move, but shift to cooperation with some probability

More non-blind cooperation maximizes payoffs.

Reputation

Larger populations: reputation is key (Nowak and Sigmund, 1998a,b).

- Cooperator or defector?
- N&S model: 'image scoring' = grading individuals on past behavior
- Individuals have scores between 0 and 5, which are updated: given an agent a with image score n at game iteration i, let a have score $n+1$ at $i+1$ if her move at i is cooperative, and $n-1$ at $i+1$ otherwise.
- Optimal strategies: always cooperate with individuals with higher score than (self|some threshold).
- Such strategies are public, so the other agent has an incentive to maintain her C-rating high: i.e. to be cooperative.

$\Longrightarrow$ Reputation enables cooperation.

Hedges . . .

- What about hedges and disclaimers?

Claim: they keep uncooperative behavior from influencing other players' expectations about the player's behavior.

- If I hedge my communication, I indicate to you not to take my move (in that game) seriously, wrt my reputation for truthfulness (/cooperativity).

To model this we need:

- Model of reputations
- Action of disclaimers within this model

After that: Issues of cooperation and restraint

A frequently expressed worry wrt this explanatory strategy:

- 'False(ish) advertising and false utterances are totally different. People hedge because they are not sure; companies put disclaimers because they want to avoid legal penalties. How can you treat these the same?'

The penalties differ, of course.

- I am claiming that the mechanism for avoiding penalties is the same and is subject to the same model.
- In both, we have acts taken to avoid damage via avoiding responsibility for some piece of content.

Both different realizations of the Jury.

Histories

Reputations derived from *histories*: sequences of objects $act \in \mathcal{A}$, $\mathcal{A}$ the set of possible actions for a given agent in a given (repeated) game.

- These objects are records of an agent's actions in past repetitions of the game.
- Game histories are n-tuples of sequences of records representing the history of the agent's actions at each decision point.
- Here: limit attention to single-move games (simplification)
  - All histories are simple sequences (1-tuples)
- Definitions simpler, life easier.

Histories for one-move games:

(4) $H_a^{g,n} = \langle act_a^1, \ldots, act_a^n \rangle$, for the action in game g by agent a, for game repetitions $1 \ldots n$.

Histories of the form above are assumed to be common knowledge of all game participants.

- Relaxing this assumption yields various refinements which I will not consider here (cf Camerer 2003).

Histories are formed by concatenation.

(5) **Making history.**
Suppose we have a move history $H_a^{g,n}$. Then $H_a^{g,n+1} = \langle act_a^1, \ldots, act_a^n, act_a^{n+1} \rangle$.

The player's latest move is simply appended to the sequence of moves preceding it.

Reputations

A player's reputation in a game is derived from his history in that game.

- A player's reputation with respect to some choice as his propensity, based on past performance, to make a particular move at that point in the game.
- Such propensities are computed from frequencies of this or that move in the history.
- Specifically, the propensity of player *a* to play a move *m* in a game *g* at move *i* is just the proportion of the total number of game repetitions that the player chose the action *m* at choice point *i*.

Frequencies for propensities are defined as follows.

$$F_{H_a^{g,n}}(move) = \frac{card(\{act \in H_a^{g,n} | act = move\})}{card(H_a^{g,n})}$$

- The above number can be viewed as a probability: in effect, the information that the game participants have about a's likelihood of choosing move m.

Example.

Suppose we have a simple game with two players, A and B, who each can make a single move, a or b, repeated three times.

- Suppose that A always plays a and B plays a twice and finally b.

Then:

(6) a. $H_A^{g,3} = \langle\langle a, a, a\rangle\rangle$

b. $H_B^{g,3} = \langle\langle a, a, b\rangle\rangle$

- A's propensity to play a is $\frac{3}{3} = 1$.
- B's propensity to play a is $\frac{2}{3} = 0.66$.

Propensities of strategies:

- 1-move games: strategies are just the frequencies above.
- Similar definition can be given for larger games by taking the product of propensities for moves.

Propensities for action

An agent's propensity to play a strategy is a real number in $[0, 1]$.

- Scalar view: an agent has a propensity for using strategy σ iff his propensity to play σ exceeds a contextually set degree s, the contextual standard for having that propensity (Kennedy, 2007). So

$$Prop(a, \sigma) \text{ iff } F_{H_a^g}\sigma \succ s,$$

where s is the contextual standard for propensity-having.

Cooperativity and truthfulness.

- A strategy σ is cooperative, $Coop(\sigma)$, if it yields a best outcome for both players over an infinite horizon of game repetitions.
- A strategy for signal-sending will be called *truthful*, $Truthful(\sigma)$, if the conventional meaning of the signal matches the sender's knowledge of the state of the world.
  - Assumes preexistence of conventional meaning mapping.
  - This can be derived by techniques in the literature also (e.g. Huttegger 2007).
- If an agent has a propensity to play a cooperative (truthful) strategy, he is said to be cooperative (truthful).

Generalizing:

- Suppose the agent plays a variety of different strategies, all of which are cooperative, but plays none of them with sufficient frequency to count as having a propensity to play that strategy.
- We still want to say the agent is cooperative, so we generalize the above to

(Note: some kind of normalization needed here: division by number of strategies?)

(7) $Coop(a)$ iff $\sum_{Coop(\sigma)} F_{H_a^g}\sigma \succ s$

(8) $Truthful(a)$ iff $\sum_{Truthful(\sigma)} F_{H_a^g}\sigma \succ s$

The situation where an agent always plays a single strategy with the relevant property is a special case.

Why cooperate?

- In the games we have seen, the best outcome for each player over a sequence of repetitions comes about if they cooperate.
- What this means is that, for repeated games, the first player has an incentive to maintain his reputation for cooperation at a high level.
- If he does not, he can expect that, at some point, the second player will stop cooperating and, again over time, the probability of a net loss will increase.

For the players, if there is a way to keep one's reputation for cooperating at a high level, there is incentive to use it.

- One way is to cooperate—the only way in games without signals.
  - (cf. the *indices* of Maynard-Smith and Harper 2003)
- But the case of games with signals, one can communicate that the signal may not be accurate.
- By sending signals of this metalinguistic kind, one can exempt oneself from damage to one's reputation, given that some other conditions are met.
- This signal is not available for application to actions—actions change the world directly and cannot be insincere or false.

Disclaimers and 'shield'-type hedges are metalinguistic signals of this type.

- They indicate: 'the signal I am about to send may be false.'
- One way to model this in standard GT: add a choice node for the first player, which indicates whether or not a disclaimer is produced.
- Instead of the frequency of the strategy of advertisement and good products, we then compute the frequency of accurate advertising with respect to the disclaimer;
- false advertising with disclaimers is thus taken to be a cooperative strategy as well.

I prefer not to take this route for two reasons.

- The disclaimer is given a character that is the same as an ordinary move. The facts seem more subtle.
- It becomes necessary to decide whether false advertising with disclaimer counts as a cooperative strategy or not.
- I do not think there is a clear answer—if there is, it is negative.
- But if it is negative, then this is approach is wrong:
- if false advertising with disclaimers are not cooperative, then disclaimers do not help to save reputations.

Better: disclaimers and hedges request that the current game not be included in the player's history.

- The disclaimer essentially says: do not hold me accountable for what I am signaling here.

This idea can be formalized in a number of different ways. Three possibilities:

1. Do not add the current game result to the stack at all—it is irrelevant to the computation of cooperativity.
2. Add the result of the game to the histories of the players, but (for the signaler) to flag it so that it does not enter the computation of cooperativity.
3. Place the 'disclaimed' signals in a separate history of their own.

Choosing depends on how players used flagged signals.

- If completely ignored in future, then the first option is right.
- If used in evaluating future signals that are also signaled as unreliable, then the third option.
- If signals that turn out to be accurate are always used, and disclaimers only are relevant to false signals, then the second or third strategy.

It seems obvious to me that the second or third options are right.

- I show one way to formalize the second and third strategies.

S2: add a second history for games that have been disclaimed.

- Each player's history is now a pair $H_a^{g,n} = \langle Hn_a^{g,n}, Hd_a^{g,n} \rangle$ of ordinary histories and the histories of disclaimed games.

(9) **Histories of the disclaimed.**

(i) Suppose we have a history $H_a^{g,n} = \langle Hn_a^{g,n}, Hd_a^{g,n} \rangle$. Then $H_a^{g,n+1} = \langle Hn_a^{g,n+1}, Hd_a^{g,n} \rangle$ if $\neg D(g_{n+1})$.
(ii) Suppose we have a history $H_a^{g,n} = \langle Hn_a^{g,n}, Hd_a^{g,n} \rangle$. Then $H_a^{g,n+1} = \langle Hn_a^{g,n}, Hd_a^{g,n+1} \rangle$ if $D(g_{n+1})$.

- When calculating truthfulness and cooperativity, we use projection functions over $H_a^{g,n}$ to check:
  - the second element of the history tuple for propensities,
  - but the first for nondisclaimed games.
- Note that we must also restate the definitions of frequency with projection functions. Details omitted.

To treat only disclaimed games with false signals as out of calculation, use (9), but modify the way that truthfulness is calculated.

- The current mechanism calculates the proportion of the total number of repetitions where the signal sent matched the state of the world.
- Here, we would like to calculate the total number of truthful signals, but ignore signals from disclaimed games only if they are inaccurate.
- We can do this with the following redefinition of frequency.

$$F^{3}_{H_a^{g,n}}(move) = \frac{card(\{a | a \in \pi_1(H_a^{g,n}) \cup \pi_2(H_a^{g,n}) \;\&\; a = move\})}{card(\pi_1(H_a^{g,n})) + card(\{a \in \pi_2(H_a^{g,n}) | a = move\}))}.$$

Which of the above strategies is right?

- I suspect that there is not a right answer about which is the correct one, and that each has its uses in particular contexts.
- This is reminiscent of interpretations of probability (cf. Binmore 2009).

Exactly what interpretations get used where is a question that would bear a lot of empirical research, which I have not done (yet?).

Hedging "compositionally"

What is the source of disclaimed readings?

- I assumed a predicate D which serves to disclaim game iterations. From where?
  $\Longrightarrow$ Do hedges have a compositional semantics?

(10) e.g. $\lambda p.D(p)$

Doubtful. What would be the source of this content? Systematic ambiguity?

- So I treat them as just metalinguistic signals.
- Claim: they arise due to charitable interpretation about incoherent discourse moves.
- ϕ together with $\Diamond\neg\phi$ Moore-inconsistent.
- Interestingly though requires parenthetical/appositive: connection with CIs ...

Suppose that a conversational agent utters $\phi \wedge \Diamond \neg \phi$.

- Inconsistent: violates precondition on assertion that the speaker must believe the content that she asserts.
  - $Bel(s, \phi)$ is required for assertion of ϕ.
- For above, precondition is $Bel(s, \phi \wedge \Diamond \neg \phi)$.
  - Then $Bel(s, \phi)$.
  - But if $Bel(s, \phi)$ then ϕ is true in all the agent's doxastically accessible worlds.
  - But $\Diamond \phi$ it is required that ϕ is epistemically possible, i.e. true in some epistemically accessible world.
  - If doxastically accessible too, then $\neg \phi$ is not consistent with the truth of $Bel(s, \phi)$.
- Result: pragmatic incoherence.

Can we eliminate this problem? Two possibilities:

1: eliminate the precondition that requires belief.

- It follows from the nature of cooperative assertion/speech act (Quality).
  - Non-eliminable with retention of cooperativity.
- Can it be forced away somehow? Yes:
  1. by modifying the sentence in such a way that belief in ϕ is not required, via sentential operator.
     - But we want to maintain ϕ as the asserted content: fail.
  2. Alternatively: hedge in the sense of disclamation.
     - What we are trying to derive . . .

2: eliminate one conjunct of the assertion itself.

- E.g. weaken it so that the content ϕ is not asserted, but proffered, cf. prejacent of epistemic modals.
- Not obvious how to do this.

A different strategy is required; but we are left with the right conjunct, $\Diamond\neg\phi$, so that is going to have to be the locus of the operation.

- Parenthetical nature of this conjunct appears to allow more flexibility (due to expressivity).
- Thus the derivation of hedges should involve parentheticality in some way.

Contrast

Disclamations appear with words that signal that a disclaiming is going to be performed: *but*, *actually*, . . .

(11) a. I might be wrong, but John is going to come tonight.
b. # I might be wrong, and John is going to come tonight.
c. # I might be wrong about this. John is going to come tonight.

Contrastive words indicate a denial of expectation(Asher, 1993; Umbach, 2004; Winterstein, 2012).

- Here: denial of expectation is arising at a pragmatic level.
- Given $\Diamond\neg\phi$, ϕ might be unexpected;
- But given the claim $\Diamond\neg\phi$, an immediately subsequent claim that ϕ is more likely to be unexpected.
- The contrastive word signals that S is aware of this.

It is strange linguistic behavior to assert something and simultaneously indicate that it might not be correct.

Explicitness

Derivation of hedging interpretations is pragmatic.

- To make the utterance sensible, it's necessary for S to indicate purposeful action.
- Without this, the principle of charity is strained beyond what it can reasonably bear.

So: *but* shows H that there is a reason for the unexpected kind of utterance being made.

- Search for a reason then gives a hedged interpretation,
- if content makes such available.

Deriving hedging

Given that intentional inconsistency has been recognized, what can be done?

- Claim: Hearer looks for a way to induce consistency by manipulating asserted content.
- This can be done by allowing *D*-interpretation, if available;
  - H looks for 'minimal disclamation', the smallest content which can be disclaimed for consistency.
  - Basically an examination and selection of literals ... (at worst, larger formulas)
  - Closely related to entailment-based rankings used by Geurts (2010) for implicature.
- When candidate is found, it is disclaimed.

McCready (2015) embeds all this in a model combining dynamic semantics and Pottsian logic for conventional implicature.

Why this bit?

Question: why is the assertion disclaimed, and not the disclaimer?

1. Since the parenthetical clause is doing the disclaiming, perhaps its content just cannot be disclaimed.
   - It does seem impossible, or even paradoxical, to simultaneously deny one's denial via the very clause performing the denial.
2. The other possibility relates to the availability of disclamation of expressive content.
   - Expressive content is exempt from disclaiming operations.
   - If this is also the case for conventionally implicated content, or even content associated directly with speech acts, then it makes sense that the parenthetical content cannot be disclaimed given that it is conventionally implicated.

Both possibilities seem plausible ...

Trust and restraint

Two major residual (pragmatic) questions.

1. In the initial state, none of the players has information about each other; otherwise put, no one has any reputation yet. Why does cooperation develop in such situations?
2. The disclaimer mechanism as defined above is powerful. It seems that it could be easily abused. Why does this not happen?

Two possible reasons for initial trust

1. An initial move of 'trust' opens the door to maximizing future benefits.

- The hearer does not know the speaker's strategy; he could be cooperative or not.
- If he is indeed trying to communicate, but she chooses not to act on his signal, cooperation is likely to fail for at least the next round (depending on strategies).
- The reason is that after the initial stage of the game, the sample size of the game history is of cardinality 1, and so whatever move the hearer makes will induce a propensity calculation yielding either 1 or 0.
- Cooperative behavior is thus a good gamble in the initial state.

2. Give a larger role to habit or convention.

- Speakers are ordinarily assumed to behave cooperatively in the sense of Gricean cooperativity.
- If conversational participants respect Quality, then we would expect truthfulness in signaling.
- Expectations about the conventions of communication lead us to gamble that whatever new agents we encounter obey these conventions.
- This gamble also leads to cooperative behavior in initial game stages.

These stories actually fall together quite neatly.

Evolution of cooperation in signaling.

Cooperation as a propensity for communicators: Grice revisited.

- This is an interpretation of cooperation on which it arises due to external factors—so it is a choice for speakers, but not one that is based on any rational considerations.
- This in turn suggests an evolutionary interpretation (Skyrms, 1996).
- On this understanding, Gricean cooperation is built into the communicating individuals as a (strong) default mechanism.

It can be stated as follows.

(12) **Cooperation by default.**
If there is no evidence to the contrary, assume the other conversational participants are (Gricean) cooperators.

If we think of cooperation as a default, we can see why it occurs in initial moves of games with disclaimers.

- Since the receiver of the signal has no evidence that the signaler ever gives false information, by default she will assume that the signaler is a cooperative agent, and so that the signal is truthful.
- She will then take the signal at its face value and perform a cooperative move.

This reading of the Gricean picture looks like the right one to me.

- It has an obvious game-theoretic realization in terms of signalling games as well, cf. Lewis (1969), etc.
- Again, this can be thought of as a special case of a larger phenomenon: perhaps cooperation is the default in general, not just in communication (Tomasello, 2008).
- Initial cooperation then an evolved mechanism leading to higher probabilities of individual payoffs.

Restraining oneself

Next issue: Disclaimers are powerful.

- In principle they can be used to hedge crazy things.
- But this does not seem to happen much.
- Speakers hedge utterances they are not fully certain about, but they do not make statements that have no credibility whatsoever, hedged for safety.
  - 'I might be wrong, but it is super hot outside—don't take your coat.'
- Advertisers may 'upgrade' but they do not blatantly lie.
  - Picture of a full roast chicken for product = single chicken nugget.

What are the constraints on disclaiming?

McCready (2008): trademarking the names of new products.

- Trademarking as a naming process with the effect of creating new kind terms called *unnatural kinds*.
- Two restrictions on trademarking:
  1. One cannot trademark an existing kind term.
  2. The second is that the name cannot be misleading. '100% All Natural Australian Beef©' is out.
- No deception!
- (Mechanism for eliminating this is different though; external third party)

Commonality: hearers do not tolerate abuse of the mechanism.

- Egregiously false signal: one which does not come close to describing the world.
  - Given the available evidence of the signaler, signal has a possibility of less than chance of being true = signaler believes probably false.
- If I hedge obviously not an honest mistake.

Generalizing:

(13) **Propriety Principle.**
Games can be disclaimed only when the signaler is engaged in honest communication.

What penalties apply to a signaler who fails to follow the Propriety Principle? Perhaps . . .

1. Game in which the violation occurs fails to be disclaimed.
   - But then no incentive to disclaim.
2. Any future attempts by the signaler to disclaim game iterations will fail.
   - = receivers will no longer trust the signaler and not allow disclaiming anymore.
3. All future attempts at disclaiming games to fail, and further all games which were successfully disclaimed in the past to be placed back on the record.
   - = rejection of disclaimers is retroactive.
   - If egregious violation is sufficiently outrageous this might be right.

Compare discussion of Serre theorem yesterday [erasures].

Biscuit conditionals

Austin (1970): a construction superficially like an indicative conditional, but with what appears to be significantly different truth conditions.

- Consequent is known to be true, but the antecedent describes a condition without which learning the consequent may not be of much interest to the hearer.

(14) There are biscuits on the sideboard if you want some. (Austin, 1970)

(15) If you care, *Independence Day* is on TV tonight.

Called 'biscuit conditionals' after Austin's example.

DeRose and Grandy (1999): biscuit conditionals as conditional assertions.

- Sentences like (14) only assert their consequents if the antecedent is true, and otherwise assert nothing.
  - ie., a special kind of speech act.
- DeR&G: this is a means of avoiding falsehood.
  - Asserting a proposition implicates content of the assertion is relevant to the hearer.
  - If it is not, then a falsehood has been implicated.
- This is a strange notion of falsehood in that it arises from metaconversational principles.
  - Very strange if I say that you have indicated something false to me by (15) because you know I don't care about *Independence Day*.
  - More so since such implicature should be defeasible.

With respect to the utility of this speech act, they say the following:

> *Such a conversational maneuver would allow you to make a warranted assertion should it be called for, while at the same time shielding you from the danger of committing the conversational misdeed of making an irrelevant assertion.* *(DeRose and Grandy 1999: 411)*

But note the similarity to our standard cases of hedging and disclaimers!

- Passage suggests that the same general mechanism is at work, even down to the use of the term *shield*.

Hedged biscuits

Natural analysis: biscuit conditionals not *really* conditionals in any sense.

- They are simply hedged assertions.
- The condtional clause functions as a shield hedge of a somewhat special type.

Advantages:

- No need to postulate a special speech act, conditional assertion.
- Avoid the problem noted by Belnap (1970): BCs judged false if consequent is false.

This analysis assimilates biscuit conditionals to hedged assertions rather than to other conditional constructions.

- Evidence: more obvious shield hedges like (16) look like perfectly adequate paraphrases of biscuit conditionals.

(16) a. I don't know if you want any, but there are biscuits on the sideboard.

b. I'm not sure you actually care about this, but *Independence Day* is on tonight.

But this cannot be the end of the story.

- What exactly is being hedged here?
- Why does the hedged content in this case differs from what is hedged in the more 'object-level' cases of hedging discussed above?

Hedging implicatures

In the cases above, what is being hedged is the implicature of relevance from the consequent.

- New observation: any Gricean implicature arising from assertion can be targeted for hedging.

(17) a. If I'm not wrong, it's raining. (Quality)
$\rightsquigarrow$ It's true that it is raining

b. If you care, *Independence Day* is on tonight. (Relevance)
$\rightsquigarrow$ You care that *ID* is on tonight

c. (*Where are you from?*) If you're interested in my nationality, I'm from the US. (Quantity)
$\rightsquigarrow$ You are not asking about my home town or planet of origin

d. If I've got the order right, John got into bed and took off his shoes. (Manner)
$\rightsquigarrow$ John got into bed before he took off his shoes

Presumably these should be analyzed the same as the biscuit cases.

Implicature or presupposition?

In the case of Quality the hedged content could also be viewed as a presupposition.

- Follow e.g. Searle (1969) and takes the truth of φ to be a precondition for asserting φ (in the guise of belief)
- View such preconditions as introducing presuppositions.
  - Not an unnatural view, actually: how else to think of these things? (Norms of assertion as presuppositional?)
- Then a presupposition is being hedged.

Is this crazy?

Not necessarily. (But maybe a little.)

- In fact we can find other cases that might be analyzable as instances of presupposition hedging.

(18) a. If John really does have a daughter, then his daughter sure never does come around.
b. If (as you claim) Burkina Faso is a monarchy, then the king doesn't have very much power.

The standard cases of presupposition binding/trapping a la van der Sandt (1992).

- Could these be hedges too?
- 'Don't take the presupposition of the consequent too seriously—just a temporary assumption.'

We get a unified analysis of a range of cases of cancellation.

- Hedges on this view are metalinguistic operators able to apply to all sorts of content.
  - truth-conditional content in the literal cases
  - 'metacontent' or nonasserted content in the biscuit-style cases

The proposal brings out the commonality of all the cases.

- In each the 'consequent' is asserted, but some side effect of that assertion is cancelled.

Can this be right? Not sure, but interesting.

Other kinds of content

Returning to advertising, consider cases where information transmission is not the main communicative goal.

- Celebrity endorsements
- Attractive images
- Even photographs of menu items, etc.

Aim: leave viewer with a positive association with the product.

- Plainly truthfulness is not a sensible requirement on such advertising.

Term: *associative advertising*.

Linguistic correlate

Expressive content: a kind of meaning representing speaker's mental state but not in a truth-conditional (at-issue) manner.

(19) a. Ouch! (Kaplan, 1999)
b. Damn, man!
c. Whoah!

(20) a. I lost my damn keys.
b. Give me the fucking money.
c. Some bum tried to talk to me on the street today.
d. Yamada-sensei-ga irasshai-masita.
Y-teacher-Nom came.Hon-Pst.Hon
'Prof Yamada came.' (and Prof. Yamada is to be respected)

Latter cases look quite close to AA.

Let us write the at-issue and expressive contents as a pair, following Potts (2005).

- A multidimensional semantics . . .

Observe that negation (and other operators) only apply to the at-issue content.

(21) a. A: I lost my damn keys.
B: That's not true.
$= \langle \neg lost(a, keys), \{damn(keyLosing)\} \rangle$

b. I didn't lose my damn keys.
$= \langle \neg lost(a, keys), \{damn(keyLosing)\} \rangle$

Hedges are similar: expressive content is not hedged.

(22) a. I might be wrong about this, but I'm pretty sure it's going to fucking rain tomorrow.
$\Longrightarrow$ speaker is in emotional state characterized by $[\![\textit{fucking}]\!]$
$\Longrightarrow D(\textit{sure}(s, \textit{rain}(\textit{tmrw})))$

b. I might be wrong about who that was out there, but it looks like that asshole is waiting for you again.
$\Longrightarrow$ speaker doesn't like referent of DP
$\Longrightarrow D(\textit{waiting}(\textit{that\_guy}))$

Again:

(23) a. $[\![(22a)]\!] = \langle D(\textit{rain}(\textit{tmrw})), \{\textit{irritated}(\textit{rain}(\textit{tmrw}))\}\rangle$

b. $[\![(22b)]\!] = \langle D(\textit{waiting}(\textit{that\_guy})), \{\textit{dislike}(\textit{that\_guy})\}\rangle$

AAs seem to work to induce *alief* in the viewer.

- Alief is a kind of mental state (Szabo Gendler, 2008b,a) arising in situations where people have what amounts to a preconscious response to certain stimuli which results in certain kinds of action or reaction.
  - fear reaction when standing on a transparent platform above a large drop (despite knowing it to be safe)
  - the disinclination of experimental subjects to eat chocolate shaped like dog feces
- Note close connection of many expressives to alief-related mental states.

Interesting analogues between pictorial and linguistic representation here ...

Back to the questions

First question already answered.

- In biscuit conditionals and related hedging constructions, the hedged content is the nonliteral content of the asserted sentence.
- Nonliteral content is content which is not at issue: implicated content but perhaps presuppositional content as well.

$\Longrightarrow$ Metalinguistic device—hedging, game disclaiming—is used to disclaim 'metalevel' content.

- Not fundamentally different from initial cases of hedging.

Second question: how to determine what is being disclaimed?

- Easy: depends on content of hedge.
  - If I hedge with 'If you care' I am hedging the Relevance implicature;
  - if I hedge with 'If you're interested in my nationality' I am hedging the Quantity implicature, . . .
- The only problematic case is that of hedges like 'If it's true', 'I might be wrong' or 'I don't know if it's true',.

It is not clear whether . . .

- the content is being hedged,
- or the implication that the content is true (via Quality)is being hedged.

But the two really go together.

- If I hedge a piece of content I also call into question its truth, and thus the Quality implicature.
- And vice versa.

Why do we need to use conditionals for hedging here—'relevance conditionals'?

- Answer: we don't!
- I showed above that other shield hedges serve the same purpose.

But why only shield hedges? Why are eg. 'might', 'I suspect that', etc impossible?

- Because we don't want to alter the content of the assertion.
- We only want to block some implicatures.
- But other kinds of hedges necessarily impact propositional content.

Needed: a way to disclaim nonasserted content.

- Assume a traditional (neo)Gricean model.
- Then each utterance U associated with four implicatures ($\phi(U)$ is content of U):

(24) a. Quality: $\mathit{True}(\phi(U))$
b. Quantity: $\mathit{Appropriate}(\mathit{gran}(\phi(U))$
c. Relevance: $\mathit{Rel}(\phi(U))$
d. Manner: . . . ?? [seems quite situation-specific]

The Quant and Rel implicatures should have definitions in terms of QUD, etc a la van Rooij (2003).

BCs then hedge one (or possibly more?) of these implicatures. Eg. (';' is dynamic conjunction)

(25) $D(Rel(on\_counter(b)));on\_counter(b)$

(26) Schema: $D(implic(\phi));\phi$

This seems the right intuition.

A problematic issue:

- Implicatures are computed by hearer (on neo-Grice model).
- How can speaker hedge content she is not responsible for?

Two options again.

1. Speaker hedges projection of hearer computations.
   - Prediction: occasional mismatches. Are there such? Not sure.
2. More radical: subsentential computation of implicatures via operator a la Chierchia (2004).
   - Would solve this problem; but the guns seem a bit heavy for the application.

My guess:

- Which implicature is being hedged should depend on default inference
- 'given my knowledge, what is the speaker most likely to be unconfident about?'
- How much of this is conventional is difficult to say at this point.

Connection to minimal revisions discussed above ...

Course summary:

- Reviewed Grice, signaling game treatment, (non)expectations of cooperativity
- Strategic communication: BM games, GS games, meaning exchange games, complexity results ...
- Applications: acknowledgements, politeness.
- Reputation model and cooperation: application to hedging and biscuit conditionals.

We hope we have . . .

- conveyed interest and effectiveness of various kinds of repeated games in linguistic analysis
- convinced you of some specific analyses
- sparked interest in this line of research!

Thank you all for coming!!

Generalizing

The reputation model can be generalized to a full picture of reasonable belief in a dynamic semantic setting.

- So far: rational belief decided on the basis of reputation = reliability.
- Reliability has been construed as ‘degree of truth-telling’.
- But ‘truthfulness’ is a property of information sources generally, not just communicative agents.
- The strategy should generalize, or at least embed in a larger picture.

Background: A standard implementation

Dynamic semantics: meanings conceived as 'context change potentials'.

- CCP = potential to alter the information that a language processor has, when that sentence is accepted.
- That information viewed as a set of 'live possibilities', where a possibility is a 'possible world' or state.
- For any claim, a state will make it true or not.
- Claims themselves (propositions) can then be thought of as sets of states, those states that verify the proposition.
- Thus an agent *a*'s information is viewed as $IS_a = \{s : a \text{ thinks } s \text{ is a live possibility}\}$.
- From this, the propositions/claims the agent believes can be derived.

Information transfer

Suppose that a conversational agent makes some claim.

- Then: the sentence uttered is interpreted by a hearer,
- and this interpretation results in a change in the hearer's information state.
- This is known as an *update*.

The basic picture:

$$\boxed{\text{Utterance} \Rightarrow \text{Interpretation} \Rightarrow \text{Update}}$$

- Update is usually modeled (in its most basic form) as set intersection:
  - for a claim of p, $IS'_a = IS \cap p$.
- These updates are quite automatic on the traditional view.

Intuition and sketch

Basic idea: virtually always update with content from any source, but only 'conditionally.'

- A new sort of dynamic semantic model moving away from 'flat' information states.
- Information states σ are now complex and consist of possibly many substates.
  - Each IS is a set of worlds (simplification),
  - ordered with a 'plausibility ranking' reflecting epistemic preferences on states.

Each substate is indexed by an index $j \in \text{Source} \cup \mathcal{A}$.

- Here Source is the set of evidence sources and $\mathcal{A}$ the set of agents.
- Constrained to only hold indices which the epistemic agent has had experience with.

This set is ordered by a total ordering:

- $\preceq_a$, where $i \prec j$ iff $P(Rel(i)) < P(Rel(j))$ for $P(Rel(i))$ the probability that source i yields reliable information.

Updates are of the form $E_i\varphi$, for E_i an operator indicating source in i-type evidence.

- A sentence $E_i\varphi$ always induces update of state σ_i.
- Utterance of φ by agent a will be represented as $E_a\varphi$.

Thus:

- At this level, update with φ always takes place —
- but this is *not* the same as coming to believe φ at a global level.

Belief merge

- Global beliefs are defined on the global state σ_T resulting from unifying all substates σ_i.
- This unification is done via a merge operation ($\Cap$):
  - all substate content survives when non-contradictory;
  - in case of conflict, information from higher-ranked source trumps lower-ranked source.
- Thus the global state almost never exhibits conflicts.

Where does the ordering come from?

So far: update of substates, substates unified via merge, merge priority determined by ordering.

- But what's the source of the ordering?
- Without that, the theory looks stipulative.

Claim: the ordering is probability-based.

- The probabilities in question are probabilities of *reliability*.
- They indicate the (perceived) likelihood that information derived from the source is correct.

These probabilities arise from two factors.

1. Initial probabilities of reliability:
   - e.g direct evidence more reliable than hearsay
   - rating of individuals qua likely truth-tellers
2. Experience with reliability of the source
   - This is, again, derived from histories.

Asher, N. (1993). *Reference to Abstract Objects in Discourse*. Kluwer, Dordrecht, Reidel.
Austin, J. (1970). Ifs and cans. In *Philosophical Papers*, pages 205–232. Oxford.
Belnap, N. (1970). Conditional assertion and restricted quantification. *Noûs*, **1**, 1–12.
Binmore, K. (2009). *Rational Decisions*. Princeton.
Camerer, C. (2003). *Behavioral Game Theory*. Princeton University Press.
Chierchia, G. (2004). Scalar implicatures, polarity phenomena, and the syntax/pragmatics interface. In A. Belletti, editor, *Structure and Beyond*, pages 39–103. Oxford.
DeRose, K. and Grandy, R. (1999). Conditional assertions and "biscuit" conditionals. *Noûs*, **33**, 405–420.
Geurts, B. (2010). *Quantity Implicatures*. Cambridge University Press.
Huttegger, S. (2007). Evolution and the explanation of meaning. *Philosophy of Science*, **74**, 1–27.
Kaplan, D. (1999). The meaning of ouch and oops: Explorations in the theory of *meaning as use*. Manuscript, UCLA.
Kennedy, C. (2007). Vagueness and gradability: The semantics of relative and absolute gradable predicates. *Linguistics and Philosophy*, **30**(1), 1–45.
Lewis, D. (1969). *Convention*. Harvard.
Maynard-Smith, J. and Harper, D. (2003). *Animal Signals*. Oxford.
McCready, E. (2008). Unnatural kinds. *Journal of Pragmatics*, **40**(10), 1817–1822.
McCready, E. (2015). *Reliability in Pragmatics*. Oxford University Press.
Nowak, M. (2006). *Evolutionary Dynamics*. Belknap Press.
Nowak, M. and Sigmund, K. (1998a). The dynamics of indirect reciprocity. *Journal of Theoretical Biology*, **194**, 561–574.
Nowak, M. and Sigmund, K. (1998b). Evolution of indirect reciprocity by image scoring. *Nature*, **393**, 573–577.
Potts, C. (2005). *The Logic of Conventional Implicatures*. Oxford University Press. Revised version of 2003 UCSC dissertation.
Prince, E., Bosk, C., and Frader, J. (1982). Hedging in physician-physician discourse. In J. di Pietro, editor, *Linguistics and the Professions*, pages 83–97. Ablex.
van der Sandt, R. (1992). Presupposition projection as anaphora resolution. *Journal of Semantics*, **9**, 333–377.
Searle, J. (1969). *Speech Acts*. Cambridge University Press, Cambridge.
Skyrms, B. (1996). *The Evolution of the Social Contract*. Cambridge.
Szabo Gendler, T. (2008a). Alief and belief. *Journal of Philosophy*, pages 634–663.
Szabo Gendler, T. (2008b). Alief in action (and reaction). *Mind and Language*, **23**, 552–585.
Tomasello, M. (2008). *The Origins of Linguistic Communication*. MIT Press.
Umbach, C. (2004). On the notion of contrast in information structure and discourse structure. *Journal of Semantics*, **21**(2), 155–175.
van Rooij, R. (2003). Quality and quantity of information exchange. *Journal of Logic, Language and Information*, **12**, 423–451.