Games with Partial Representations: An Odycceus Workshop
Chalmers University of Technology, Gothenburg, September 11-12, 2017

Non-Equilibrium Strategic Thinking and the Role of Abstract or Natural-Language Communication

Vincent P. Crawford

University of Oxford, All Souls College,
and University of California, San Diego

UC San Diego

European Research Council
Established by the European Commission

I owe thanks to many friends, students, and co-authors for valuable conversations and advice on this topic. My research received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement no. 339179. The contents reflect only my views and not the views of the ERC or the European Commission, and the European Union is not liable for any use that may be made of the information contained therein. The University of Oxford, All Souls College, and the University of California, San Diego also provided funding.

Revised 13 September 2017

Introduction

The conference's main focus is on equilibrium or evolutionary analyses of games with partial representations, with particular attention to the role of communication, which plays a central role in most human interactions.

By contrast, my focus will be on non-equilibrium analyses of the role of communication in fully represented games.

Such analyses reflect people's cognitive constraints in thinking about others' behavior, much as analyses of games with partial representations reflect people's cognitive constraints in thinking about games.

Ultimately we will need models with both kinds of constraint, so I hope my focus will be complementary to the conference's main focus.

I will consider two leading examples, in which existing models seem to fall short of how people actually use communication, in different ways:

- bringing about efficient coordination in one-shot Stag Hunt games
- bringing about and maintaining efficient cooperation and coordination in long-term relationships

My discussion builds on my paper, “New Directions for Modelling Strategic Behavior: Game-Theoretic Models of Communication, Coordination, and Cooperation in Economic Relationships,” *JEP* 2016: https://www.aeaweb.org/articles?id=10.1257/jep.30.4.131

and my Nancy Schwartz Lecture, "Modeling Strategic Communication: From Rendezvous and Reassurance to Trickery and Puffery": http://www.kellogg.northwestern.edu/news-events/lecture/schwartz.aspx, slides at: http://econweb.ucsd.edu/~v2crawford/SchwartzLecture17.pdf

In each case I will assume for simplicity that players know the game's structure as *common knowledge*, so that they face uncertainty only about other players’ decisions, *strategic uncertainty*.

I will argue that:

- Our models of communication should avoid reflexively assuming equilibrium in applications where its learning and strategic thinking justifications are implausible, and try to do fuller justice to strategic uncertainty

- We should seek evidence-based models with the generality of equilibrium models, which reflect that communication is an all-purpose tool, used to deceive, persuade, reassure, or coordinate (Rendezvous, Reassurance, Trickery, and Puffery in my Schwartz Lecture's subtitle)

- We should seek models that imply substantive roles for communication (unlike most equilibrium-based models), and that can reflect the advantages of natural-language over abstract communication (even in games of complete information, where there is nothing concrete to communicate but players' beliefs and intentions)

Equilibrium as a behavioral model

Equilibrium normally requires coordination of players' beliefs about how the game will be played, which has two alternative justifications:

- *learning* from extensive experience with analogous games, which has a strong tendency to make players' beliefs converge to equilibrium

- *strategic thinking*, which can yield equilibrium beliefs even in players' initial responses to a game, under strong assumptions regarding players' knowledge of each other's beliefs (Brandenburger *JEP* 1992)

(Farrell *EL* 1988 showed that even unlimited pre-play communication does not assure equilibrium in the underlying game without the question-begging assumption of equilibrium in game with communication.)

In many applications, players don't have enough precedents to learn an equilibrium, and the thinking justification is too complex to be plausible.

In experiments that study learning in repeated play of identical stage games with different partners, subjects usually converge to some stage-game equilibrium; but their designs (like most of our theories of learning) overstate the clarity of precedents in applications.

Experiments that elicit initial responses to games suggest that people's thinking seldom follows the fixed-point or indefinitely iterated dominance reasoning that equilibrium requires in all but the simplest games.

Yet in economically interesting games, replacing equilibrium with weaker assumptions like *rationalizability* (Bernheim *Ecma* 1984 or Pearce *Ecma* 1984) yields few or no restrictions on behavior.

I will suggest that *level*-k or *cognitive hierarchy* models are an evidence-based first step toward relaxing equilibrium, while retaining the power and generality of equilibrium analysis; but stop well short of the theory we need to understand how people use communication in relationships.

One-shot Stag Hunt games

In Rousseau's example (*Discourse on Inequality* 1754 [1973]):

> If a deer was to be taken, everyone saw that, in order to succeed, he must abide faithfully by his post: but if a hare happened to come within the reach of any one of them, it is not to be doubted that he pursued it without scruple, and, having seized his prey, cared very little, if by so doing he caused his companions to miss theirs.

	Stag	**Hare**
Stag	**9** **9**	**8** **0**
Hare	**0** **8**	**7** **7**

Stag Hunt

(Here I follow Aumann's 1990 and Charness's *GEB* 2000 payoffs.)

The main issue is the tension between the higher potential payoff of playing Stag and its greater strategic uncertainty.

Equilibrium in one-shot Stag Hunt without communication

	Stag	Hare
Stag	9 9	8 0
Hare	0 8	7 7

Stag Hunt

Stag Hunt without communication has three equilibria:

- Two symmetric pure-strategy equilibria: “both-Stag” and “both-Hare”
- One symmetric mixed-strategy equilibrium

The mixed-strategy equilibrium is arguably behaviorally irrelevant in this game, and I will ignore it.

	Stag	Hare
Stag	**9** **9**	**8** **0**
Hare	**0** **8**	**7** **7**

Stag Hunt

Both-Stag is better for both players than both-Hare (or the mixed-strategy equilibrium): “payoff-dominant” (Harsanyi and Selten 1988).

But both-Hare has players choosing best responses to a larger range of players’ beliefs: “risk-dominant”.

Harsanyi and Selten favor payoff-dominance over risk-dominance, and therefore favor both-Hare over both-Stag or the mixed equilibrium.

Yet in experiments that elicit initial responses to such games, most subjects play risk-dominant strategies like Hare, but a few play payoff-dominant strategies like Stag, and so the aggregate choice frequencies deviate systematically from those of either pure or the mixed equilibrium.

Level-*k* thinking as a behavioral model

Much evidence from experiments that elicit initial responses to games without communication points to a class of non-equilibrium models based on level-*k* thinking (Stahl and Wilson *JEBO* 1994, *GEB* 1995; Nagel *AER* 1995; Costa-Gomes, Crawford, and Broseta *Ecma* 2001; Camerer et al. *QJE* 2004 (*cognitive hierarchy* models); Costa-Gomes and Crawford *AER* 2006; surveyed in Crawford, Costa-Gomes, and Iriberri *JEL* 2013).

In a level-*k* model, players follow rules of thumb that:

- anchor their beliefs in a naïve model of others' responses, called *L0*

and

- adjust their beliefs via a small, heterogeneous number (*k*) of iterated best responses: *L1* best responds to *L0*, *L2* to *L1*, and so on

In games without communication, most of the evidence is consistent with *L0* being uniform random over the feasible decisions.

This *random L0* reflects higher levels' initial thinking about the incentives the payoff structure creates, using the principle of insufficient reason as a proxy for others' behavior before they consider others' responses to incentives (Crawford et al. *JEL* 2013, especially Section 4).

The population frequencies of higher levels are treated as behavioral parameters.

The frequency of *L0* if freely estimated is usually zero or small, and the estimated level distribution is normally concentrated on *L1*, *L2*, and *L3*.

Importantly, a level-*k* model makes precise (probabilistic) predictions: not only that deviations from equilibrium will sometimes occur, but also which settings evoke them and which forms they are likely to take.

- Level-*k* players use step-by-step procedures that generically determine unique pure strategies, with no need for fixed-point reasoning or indefinitely iterated dominance

- *Lk* (for $k > 0$) is decision-theoretically rational, with an accurate model of the game; it departs from equilibrium only in deriving its beliefs from an oversimplified nonequilibrium model of others' responses

- *Lk* (for $k > 0$) respects *k*-rationalizability (Bernheim 1984 *Ecma*), hence in two-person games its choices survive *k* rounds of iterated elimination of strictly dominated strategies

- Thus *Lk* (for $k > 0$) mimics equilibrium decisions in *k*-dominance-solvable games, but can deviate systematically in other games

  (Such deviations make it possible for a level-*k* model to systematically out-predict a rational-expectations notion such as equilibrium)

- A level-*k* model with zero weight on *L0* can thus be viewed as a heterogeneity-tolerant refinement of *k*-rationalizability
  (even without *L0*, a cognitive hierarchy model cannot be so viewed)

Level-*k* thinking in one-shot Stag Hunt without communication

	Stag	Hare
Stag	9 9	8 0
Hare	0 8	7 7

Stag Hunt

When *L0* is random *L1* plays Hare, and all higher levels do so too: Thus a level-*k* model predicts the same outcome as risk-dominant equilibrium.

In this case a level-*k* model has no advantage in fit over the best refined-equilibrium model.

However, that equilibrium and refinements play no role in players' thinking makes its prediction here cognitively more plausible.

And a level-*k* model has a significant advantage in fit in other settings (Crawford et al. *JEL* 2013, especially Section 3).

Modeling communication

I focus on games in which communication takes the form of players sending one or more rounds of one- or two-sided pre-play messages.

In equilibrium analyses of games without private information, messages must be about players' *intentions* about their choices in the *underlying game* (Kalai and Samet *IJGT* 1985, Farrell *Rand* 1987, Rabin *JET* 1994).

More generally, messages might also concern *private information* (Crawford and Sobel *Ecma* 1982, Green and Stokey *JET* 2007 [1981]).

I will assume that messages are *cheap talk*, in that they have no direct effect on payoffs; thus messages about intentions must be nonbinding.

If cheap-talk messages are in a pre-existing common language that makes lying a meaningful concept, lying must have no direct cost (a useful case even if people are actually averse to lying; Ellingsen and Johannesson *EJ* 2004 and Abeler, Nosenzo, and Raymond 2016).

Subgame-perfect equilibrium with communication as a behavioral model

Subgame-perfect (or *sequential* or *perfect Bayesian*) equilibrium rules out equilibria that do not prescribe equilibrium play in subgames (etc.).

Even so, subgame-perfect equilibrium predictions with cheap talk are ambiguous for other reasons. I focus on *sensible* (my term) equilibria, which rule out behaviorally unimportant ambiguities (but leave in others):

- There is always a *babbling* equilibrium in which messages are uninformative; but I focus on informative equilibria when they exist

- Payoffs-based refinements cannot determine the literal meanings of cheap-talk messages, so I focus on equilibria in which literal meanings are understood ("Yes means Yes"), whether or not they are believed (Similarly, level-*k* models anchor beliefs on *L0*s that respect meanings)

In a sensible equilibrium, the operative meaning of a cheap talk message is "I like what I expect you to do when I say this, better than anything I could induce you to do by saying something else."

Subgame-perfect equilibrium in one-shot Stag Hunt with communication

	Stag	Hare
Stag	9 9	8 0
Hare	0 8	7 7

Stag Hunt

- In Stag Hunt with one round of one-sided communication, there are two sensible equilibria, one in which the sender sends and plays Stag, and another in which the sender sends and plays Hare

- With one round of two-sided communication, there are again two sensible equilibria, one in which both players send and play Stag, and another in which both send and play Hare

- With many rounds of two-sided communication there are again multiple sensible equilibria, which do not improve upon those with one round

Payoffs-based refinements cannot (by definition) determine the operative meanings of cheap-talk messages, but Farrell *EL* 1988, *GEB* 1993; Myerson *JET* 1989; Rabin *JET* 1990; and Farrell and Rabin *JEP* 1996 have proposed language-based refinements that can do so:

- A *self-signaling* message regarding private information or intentions is one that a sender wants a receiver to believe if and only if it's true

- A *self-committing* message regarding intentions (only) is one that, if believed, creates an incentive for the sender to do as s/he said; that is, makes the sent intention part of an equilibrium in the underlying game

Aumann 1990 notes that in Stag Hunt, a one-sided message of "Stag" is self-committing but not self-signaling.

On that basis he argues that such a message can convey no information and therefore cannot affect the outcome in the underlying game (Farrell *EL* 1988, Rabin *JET* 1994 disagree; see also Crawford *JEP* 2016).

Yet in experiments, messages of “Stag” often yield coordination on both-Stag (Cooper, DeJong, Forsythe, Ross *QJE* 1992; Charness *GEB* 2000; Duffy and Feltovich *GEB* 2002; but see Clark, Kay, Sefton *IJGT* 2001; Dugar and Shahriar 2016; Ellingsen, Östling, Wengström 2016).

Two-sided messages do as well as one-sided, but not significantly better.

Natural-language messages, one-sided, single-round or with dialogues, do much better (Dugar and Shahriar 2016)

Why does behavior in experiments deviate so much, so systematically, from Aumann's prediction? Two conjectures:

- Aumann assumes that players' beliefs are focused on a particular equilibrium, even though strategic uncertainty is the essence of the problem. Few people will assume that even a message that is not self-signaling will have no influence on others' choices; and the assumption that it will have no influence goes far beyond rationality

- Aumann's analysis implicitly limits players to a fixed list of messages of Intent when they would plainly benefit from a more nuanced discussion

Farrell *EL* 1988 and Rabin *JET* 1994 (see also Myerson *Ecma* 1983, *JET* 1989) follow the first route by relaxing equilibrium to rationalizability with behavioral restrictions on how players use language.

They get strong results on the effectiveness of communication in Stag Hunt, which predict the efficient equilibrium outcome; but their methods yield weaker predictions when communication serves other purposes.

Level-*k* thinking with communication as a behavioral model

The goal is to identify a model of communication with the generality of equilibrium models, which yields precise and reliable predictions, up to behavioral parameters, across games where communication serves a variety of purposes (Rendezvous, Trickery, Puffery, Reassurance).

I now adapt the level-*k* model to allow communication, as in Crawford *AER* 2003 (see also Cai and Wang *GEB* 2006; Kartik, Ottaviani, Squintani *JET* 2007; and Ellingsen and Östling *AER* 2010; "EÖ").

Senders' beliefs are anchored in *L0*s that favor the truth, as in Crawford *AER* 2003 (who considered only one-sided messages about intentions); but for two-sided messages *L0* receivers randomize uniformly independent of received messages, as in EÖ.

The rest of the model is specified by iterating best responses as before.

The resulting model is well supported by experimental evidence (Cai and Wang *GEB* 2006; Kawagoe and Tazikawa *GEB* 2009; EÖ; Wang, Spezio, Camerer *AER* 2010, Dugar and Shahriar 2016, Ellingsen, Östling, Wengström 2016, and García-Pola, Iriberri, Kovářík 2016).

Level-*k* thinking in one-shot Stag Hunt with communication

	Stag	Hare
Stag	9 9	8 0
Hare	0 8	7 7

Stag Hunt

- With one round of *one*-sided communication, *L1* senders send and play their risk-dominant strategy, Hare. *L1* receivers play Hare when they hear Hare and Stag when they hear Stag. *L2* and higher senders send and play Stag, expecting to be believed; and *L2* and higher receivers, expecting to hear and play Hare, play Stag when they hear Stag, which is self-committing for them. Thus the level-*k* model predicts that if both players are *L2* or higher, they will coordinate on both-Stag
- With one round of *two*-sided communication, *L2* and higher players again send and/or play Stag, as do *L1* players with positive probability in EÖ's model, making two-sided communication more effective
- Multiple rounds of two-sided communication are no better than one

These predictions come closer than refined equilibrium notions to fact patterns in experiments with Stag Hunt and, importantly, other games.

But intuitively, natural-language dialogues would make coordination on the efficient both-Stag far more likely. A Sender could begin as follows:

"I can see, as I'm sure you can, that the best possible outcome would be for both of us to play Stag. I realize that Stag is risky for you, as it is for me. But despite the risk, I think Stag's higher potential payoff makes it a better bet. I therefore plan to play Stag, and I hope you will too."

To my knowledge only Dugar and Shahriar 2016 have done natural-language experiments with Stag Hunt, but such messages are likely to yield efficient coordination at a higher rate than abstract dialogues.

Yet we have no model that can represent possible differences between natural-language and abstract messages (partial exceptions: Selten and Warglien *PNAS* 2007; Hong, Lim, Zhao 2016; Gibbons, LiCalzi, Warglien 2017). (Even though natural language stories, viewed as internal monologues, not actual dialogues, motivate equilibrium refinements.)

Long-term relationships

A central question in economics is how people bring about and maintain efficient cooperation and coordination in a long-term relationship.

Most analyses of long-term relationships assume that players play a particular subgame-perfect (etc.) equilibrium of the repeated game that describes the entire relationship, and then characterize the Folk Theorem set of outcomes consistent with some such equilibrium. Yet:

- The size of the set of equilibria, the idiosyncrasies of relationships, and the fact that we don't get to practice them make equilibrium via learning or thinking behaviorally implausible

- Such analyses seldom consider robustness to strategic uncertainty, and focus for convenience on equilibria that are "brittle" (but see Porter *JET* 1983, van Damme *JET* 1989, or Friedman and Samuelson 1994)

- Further, in an equilibrium of a repeated complete-information game, players have nothing concrete to communicate, despite communication's powerful influence in practice

Imagine that you are in a long-term relationship, governed by an implicit agreement you believe you and your partner understand. For the first time, your partner doesn't do what you thought was agreed. What now?

Some observations and intuitions:

- Without communication, all you can do is signal your displeasure via tit-for-tat, hoping your partner will understand and return to cooperation. If the ideal agreement is as obvious as in models, such tactics might work (van Huyck, Battalio, and Beil's *AER* 1990 fixed-pair treatments).
- However, even though most equilibrium-based analyses theoretically make tacit collusion a perfect substitute for explicit collusion, if good agreements require non-obvious choices or non-obvious surplus-sharing, restoring cooperation is hopeless without communication.

  That's why antitrust law bothers to prohibit firms from communicating (Genesove & Mullin *AER* 2001, Andersson & Wengström *SJE* 2007).
- With abstract communication via a set list of understood messages, you might be able to restore cooperation, but only if good agreements are simple and reaching them doesn't require complex adjustments.

- By contrast, with natural-language messages, single or dialogue, there is hope to restore cooperation:
  - If only a single, one-sided message is possible, a contingent promise to return to cooperation if your partner does so might work
  - Even if identifying a good agreement is complex, restoring cooperation may be possible via a natural-language dialogue, e.g. starting like this:

    "I value our relationship, and I believe you are trying to cooperate. But what you just did was inconsistent with what I thought we had agreed. [Elaborates….] Please help me to understand your thinking"

A growing body of evidence suggests that natural-language dialogues are far more effective than structured abstract communication (Valley, Thompson, Gibbons, and Bazerman *GEB* 2002; McGinn, Thompson, and Bazerman *JBDM* 2003; Charness and Dufwenberg *Ecma* 2006, *EL* 2010; Cooper and Kühn *AEJ Micro* 2014; Dugar and Shahriar 2016).

Yet again we have no model to represent differences between natural-language and abstract messages (partial exceptions: Selten & Warglien *PNAS* 2007; Hong, Lim, & Zhao 2016; Gibbons, LiCalzi, Warglien 2017).

Future work

My examples highlight large gaps between intuition, evidence, and theory about how people use communication to structure relationships.

Three lines of experimental, empirical, and theoretical research seem likely to be especially helpful:

- Work on strategic thinking and behavior in games without communication, particularly with nontrivial sequential structures

  Recent examples include Dal Bó and Fréchette *AER* 2011; Blonski, Ockenfels, and Spagnolo *AEJ Micro* 2011; Ho and Su *MS* 2012; Kawagoe and Takizawa *JEBO* 2012; Breitmoser *AER* 2015; and García-Pola, Iriberri, and Kovarik 2016.

- Work explaining why and how communication (abstract or natural-language) allows people to achieve (behaviorally, not theoretically) outcomes better than those attainable without communication

  Examples, following Myerson *Ecma* 1983, *JET* 1989 and Forges *Ecma* 1986, include Weber and Camerer *MS* 2003; Houser and Xiao *EE* 2010; Andersson and Wengström *JEBO* 2012; Cooper and Kühn *AEJ Micro* 2014; Awaya and Krishna *AER* 2016; Dugar and Shahriar 2016)

- Finally and most challengingly, work explaining why and how natural-language communication, particularly in unlimited dialogues, improves upon structured communication via abstract messages

  The message I suggested above for Stag Hunt shows that even a single natural-language message can convey an understanding of strategic issues, essential in some settings, that can theoretically but not *behaviorally* be conveyed via abstract messages.

  In theory players can mentally simulate any natural-language message or dialogue (Myerson *Ecma* 1983, *JET* 1989), but in practice that is no substitute for actual communication; Myerson *Ecma* 1983, *JET* 1989 and Forges *Ecma* 1986 model messages more richly than usual.

  A further puzzle is why dialogues are better than “brief-filing”: they economize on cognition and bandwidth, and Forges *Ecma* 1986 and Myerson *Ecma* 1986 show they may expand possibilities other ways.

  There is very little further work on this topic, considering its importance; recent examples include Genesove and Mullin *AER* 2001; Cooper and Kagel *AER* 2005; Charness and Dufwenberg *Ecma* 2006, *EL* 2010; Cooper and Kühn *AEJ Micro* 2014; Burchardi and Penczynski *GEB* 2014; Awaya and Krishna *AER* 2016; and Dugar and Shahriar 2016.

I note in closing that in studying cognition, it is likely to be helpful to take fuller advantage of experimental methods that measure it more directly:

- Monitoring subjects' searches for information about payoffs (Costa-Gomes et al. *Ecma* 2001; Johnson, Camerer, Rymon, and Sen *JET* 2002; Costa-Gomes and Crawford *AER* 2006; Wang et al. *AER* 2010; Brocas, Carillo, Wang, and Camerer *REStud* 2014)

  (The earlier work is surveyed in Crawford 2008 http://econweb.ucsd.edu/%7Evcrawfor/5Oct06NYUCognitionSearchMain.pdf)

- Monitoring the chats of teams of subjects who must agree on decisions before they are implemented (Moreno and Wooders *GEB* 1998; Cooper and Kagel *AER* 2005; Burchardi and Penczynski *GEB* 2014)