Deception as Cooperation

Abstract

I develop a rate-distortion analysis of signaling games with imperfect common interest. Sender and receiver should be seen as jointly managing a communication channel with the objective of minimizing two independent distortion measures. I use this analysis to identify a problem with 'functional' theories of deception, and in particular Brian Skyrms's: there are perfectly cooperative, non-exploitative instances of channel management that come out as manipulative and deceptive according to those theories.

1 Introduction

How communication is modeled in a Lewis-Skyrms signaling game (also simply *signaling game* henceforth, Lewis 1969, chap. 4; Skyrms 2010) is perfectly isomorphic to how information processing is modeled in information theory (Shannon & Weaver 1998; Cover & Thomas 2006). See Figs. 1 and 2.

Figure 1: A signaling game

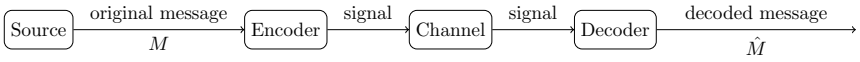

Figure 2: An information-processing pipeline

In this paper I take this isomorphism seriously: literally, senders and receivers in a signaling game are jointly managing an information-processing channel. This information-theoretic perspective on signaling games has been neglected in the literature, and it should not have: as I will argue here, the central behavioral unit in signaling games is neither sender nor receiver strategies, but the encoding-decoding pair (also *code* henceforth) that they jointly construct. This 'channel-first' perspective makes it possible to identify cooperative examples of joint channel-management. It will turn out that what certain prominent

contemporary accounts of deception call 'deceptive' or even 'manipulative' is compatible with exquisitely cooperative arrangements.

I have just suggested that information-theoretic analyses of signaling games are virtually non-existent. If the reader is familiar with the literature, this claim might have struck them as odd: information-theoretic notions *are* routinely used in the analysis of signaling games. Unfortunately, only the subset of information theory that was introduced to philosophers by Dretske's seminal (1981) book is typically appealed to. This subset is relatively small, and can be introduced in full in a few paragraphs. I will do so in section 2, after briefly presenting signaling games. The Dretskean subset of information theory leaves important informational structure out. In section 3 I summarize a few key ideas in *rate-distortion theory*, the branch of information theory that describes lossy communication and which, I propose, can be fruitfully applied to the description of signaling games. In section 4 I discuss a conservative extension to rate-distortion theory that deals with situations in which two independent distortion measures (in our case, those of sender and receiver) are used to calculate the rate of a code. With the help of these tools, in section 5 I show that a very prominent, so-called *functional* approach to deception in simple organisms, and in particular Skyrms's application of this idea, somewhat myopically regards individual signals as manipulative, even if they are part of a code which equally respects the interests of sender and receiver. Section 6 recapitulates and offers some conclusions.

2 Information in Lewis-Skyrms Signaling Games

In the version of signaling games I will be concerning myself with here, the world observed by the sender is represented as a *random variable*, S.[1] For my current purposes, this random variable can be identified with a set of s possible states for the variable to be in, $[S_1, \ldots, S_i, \ldots, S_s]$, together with a probability distribution over states $[\Pr(S_1), \ldots, \Pr(S_s)]$, where $\Pr(S_i) \geq 0$ and $\sum_i \Pr(S_i) = 1$ (MacKay 2003, p. 34) The sender observes the state the world is in and then sends a signal to the receiver (see Fig. 1). Again, signals can be thought of as a random variable M: there are, say, m possible signals in the putative representation system we are studying, $[M_1, \ldots, M_i, \ldots, M_m]$. Finally, the receiver observes the signal sent by the sender and chooses an act, A_i, out of a possible acts. A is our third and final random variable.

There is a payoff associated with every combination of state and act.[2] The combination of state S_i and act A_j results in a sender payoff of p^{σ}_{ij} and a receiver payoff of p^{ρ}_{ij} (Fig. 3). Signaling games are, as originally described,

[1] For a more detailed introduction to Lewis-Skyrms signaling games see Martínez & Godfrey-Smith (2016).

[2] Here I am focusing on so-called *cheap talk* games. In related models, payoffs attach to triples of state, message and act. See, e.g., Martínez (2015).

game-theoretic models, and these payoff matrices are used to calculate Nash equilibria and evolutionary trajectories (see Godfrey-Smith & Martínez 2013; Martínez & Godfrey-Smith 2016 for details). As we are about to see, that is the full extent of their use—again, it should not be.

$$\begin{bmatrix} p^{\sigma}_{11} & \dots & p^{\sigma}_{1a} \\ \vdots & \ddots & \vdots \\ p^{\sigma}_{s1} & \dots & p^{\sigma}_{sa} \end{bmatrix} \qquad \begin{bmatrix} p^{\rho}_{11} & \dots & p^{\rho}_{1a} \\ \vdots & \ddots & \vdots \\ p^{\rho}_{s1} & \dots & p^{\rho}_{sa} \end{bmatrix}$$

Figure 3: Sender payoff (left) and receiver payoff (right)

The probabilities associated with the three random variables, S, M and A, are interconnected through the *sender strategy* (a matrix of probabilities of signals conditional on world states), and the *receiver strategy* (a matrix of probabilities of acts conditional on signals). See Fig. 4.

$$\begin{bmatrix} \Pr(M_1|S_1) & \dots & \Pr(M_m|S_1) \\ \vdots & \ddots & \vdots \\ \Pr(M_1|S_s) & \dots & \Pr(M_m|S_s) \end{bmatrix} \begin{bmatrix} \Pr(A_1|M_1) & \dots & \Pr(A_a|M_1) \\ \vdots & \ddots & \vdots \\ \Pr(A_1|M_m) & \dots & \Pr(A_a|M_m) \end{bmatrix}$$

Figure 4: Sender strategy (left) and receiver strategy (right)

With the probabilities of world states together with sender and receiver strategies we can calculate all possible joint and conditional probabilities involving the three random variables. Brian Skyrms, in a very influential information-theoretic treatment of signaling games (2010, ch. 3), uses these quantities to define the *informational content* of a signal. First, he defines the "information in $[M_j]$ in favor of $[S_i]$" (op. cit., p. 36—variables changed) as

$$\log_2 \Pr(S_i|M_j) - \log_2 \Pr(S_i) \tag{1}$$

This is simply a measure of the difference between the unconditional probability of world state S_i and its probability conditional on the signal.[3] Skyrms then proposes that the informational content of a signal, M_i, should be identified with the collection of the informations that this signal carries about every world state:

$$\langle \log_2 \Pr(S_1|M_i) - \log_2 \Pr(S_1), \dots, \log_2 \Pr(S_s|M_i) - \log_2 \Pr(S_s) \rangle \tag{2}$$

The analogous construction gives us the informational content of the signal about acts (*ibid.* p. 39).

[3]This quantity is sometimes called *pointwise mutual information*, and is widely used in automatic text parsing (e.g., Bouma 2009).

In the information-theoretic constructions in his 2010 book and, as far as I am aware, elsewhere, Skyrms uses *all of* but also *only* three mathematical objects: the probability distribution of states, the sender strategy, and the receiver strategy. Both the "all of" and the "only" are regrettable. The "all of" means, among other things, that no informational analysis is possible until the sender and the receiver have made up their mind as to what strategy to adopt. Postponing informational analyses until a sender and a receiver strategy are available tacitly assumes that information theory can offer no guidance as to *what has led sender and receiver to adopt those very strategies*.[4] Typically, indeed, prominent sender-receiver configurations (e.g., Nash equilibria, or sinks in evolutionary dynamics) are singled out for information-theoretic investigation. The underlying idea seems to be that only (evolutionary) game-theoretic properties of the signaling setup are relevant to sender and receiver adopting the strategies they do adopt. Such features are, of course, *sufficient* for those strategies to obtain—this is just the way signaling games are designed, after all. On the other hand, sender and receiver are solving an *informational* problem: that of what to communicate, and how much to communicate about it, given their interests. This is also by design: signaling games are precisely a tool formulated to study communication. It should therefore be possible to identify informational constraints on available, and attractive, strategies for sender and receiver to adopt. Yet the mainstream information-theoretical toolbox in philosophy, the one used by Skyrms and other following on Dretske's (1981) footsteps, has no resources for making information theory contribute to our understanding of the informational structure of a signaling gane, independently from this or that sender-receiver configuration. That is, it has no way to answer the question: what are the informational properties of the game setup that help explain players having ended in this or that sender-receiver configurations?

Regarding the "only", it is surprising that payoff matrices should play no role in the informational analysis of signaling games. As I suggested above, sender and receiver have possibly different desiderata as to what to communicate. Plausibly these sets of interests will influence the properties of the code that their strategies jointly constitute—what it singles out for transmission, and what it discards. These sets of interests are given by the payoff matrices, which should therefore be factored in our informational analysis.

These shortcomings are a result of the undue focus on *the information carried by particular vehicles* in most philosophical treatments of information (Kraemer 2015; Scarantino 2015; Shea 2007; Skyrms 2010; Stegmann 2015, among many others). In the Lewis-Skyrms framework, this translates into a focus on the informational properties of particular signals, and in particular of the information

[4] As an anonymous referee helpfully remarked, this is not to say that Skyrms's overall theory does not offer such guidance. It does: it is the evolutionary game-theoretic aspect of his theory that deals with the evolution of sender-receiver strategies—including the evolution of the informational properties of these strategies (see, e.g., Skyrms 2010, p. 40, fig. 3.3).

The information-theoretic ingredient of Skyrms's theory is silent about this, and only pertains to the description of static, instantaneous sender-receiver configurations.

they carry about states, on the one hand, and acts, on the other. This piecemeal, one-sided evaluation of the informational properties of signals obscures the role they play in the *mediation* between states and acts. It is this mediation that signals are for.

3 Back to Shannon

Information theory proper, beyond the fragment Dretske chose to focus on and interpret, offers a better analysis of signaling games. The main aim of information theory, the branch of applied mathematics inaugurated by Claude E. Shannon's astonishing *The Mathematical Theory of Communication* (1948; the standard textbook treatment is Cover & Thomas 2006) is to identify *theoretical bounds for distortion in the transmission of information through a (typically noisy, typically narrow) channel*. Signaling games can be described with the tools of information theory because they are just that: information-processing pipelines, in which a channel and a distortion measure can be readily identified.

First, as regards the channel, sender and receiver strategies in a signaling game are, quite literally, an encoder-decoder pair (again, see Figs. 1 and 2): the sender transforms world states incoming from the source into signals. The receiver decodes those signals to obtain acts. The matrix that results from multiplying sender and receiver strategies (Fig. 5), and that transforms states into acts, identifies a rate equal to the mutual information between states and acts, $R = I(S;A)$. By the *source-channel separation theorem with distortion* (Cover & Thomas 2006, Theorem 10.4.1) the channel through which signals are sent must have a capacity, C, such that $C > R$.

$$\begin{bmatrix} \Pr(A_1|S_1) & \dots & \Pr(A_a|S_1) \\ \vdots & \ddots & \vdots \\ \Pr(A_1|S_s) & \dots & \Pr(A_a|S_s) \end{bmatrix}$$

Figure 5: A Lewis-Skyrms code

Second, signaling games come equipped with a measure of distortion in the transmission of information: the payoff matrices provide precisely an answer to the question, how good is decoding state S_i as A_j, compared to the best we could possibly do? This observation can be turned into a formal distortion measure simply by normalizing and rescaling the payoff matrices.[5] Setting $p_{max} = \max_{ij}(p_{ij})$, $p_{min} = \min_{ij}(p_{ij})$, the distortion measure for the pair (S_i, A_j) is

[5] One could carry out all of the analyses in this paper, *mutatis mutandis*, using payoff values directly. I will transform them into normalized distortion measures, though, as is standard in rate-distortion analyses.

$$d_{ij} = \frac{p_{max} - p_{ij}}{p_{max} - p_{min}} \tag{3}$$

Once we have a channel and a distortion measure, one central result in information theory is that it is possible to calculate a *rate-distortion* function (Shannon 1948, 1959; Cover & Thomas 2006, ch. 10) that gives the minimum rate, and hence channel capacity, sufficient to achieve any expected level of distortion D.[6] This function is given by:

$$R(D) = \min_{p(a|s):\sum_{(s,a)} p(s,a)d(s,a) \leq D} I(S;A) \tag{4}$$

That is, it's given by the minimum mutual information between states and acts that still meets the distortion goal: "the minimization is over all conditional distributions $\Pr(A|S)$ for which the joint distribution $\Pr(S, A)$ satisfies the expected distortion constraint [D]" (Cover & Thomas 2006, p. 335—variables changed). This minimization, in sum, gives the minimum rate at which a certain distortion, or less, is achievable. The Blahut-Arimoto algorithm (Blahut 1972; Arimoto 1972; Cover & Thomas 2006, sec. 10.8) provides an efficient way to calculate the $R(D)$ function.

The rate-distortion function depends only, on the one hand, on the probabilistic structure of the source, S, and, on the other hand, on the distortion measure D. Both of these quantities are prior to, and do not depend, on the actual strategies being implemented by sender and receiver. The rate-distortion function, thus, offers a way to characterize the informational structure of a signaling game, independently of the actual behavior of senders of receivers—indeed, I will be defending in what follows that it is illuminating to see senders and receivers as *reacting* to this informational landscape.

For a first example of how these rate-distortion functions look like consider one of the simplest signaling games: the 3x3 Lewis signaling game.[7] In this game there are three equiprobable states and three possible receiver acts. The payoff for every combination of state and act is given by Table 1. Table 2 shows the distortion measure that corresponds to this payoff matrix (i.e, the payoff matrix rescaled and normalized as per eq. 3). The rate-distortion curve for this Lewis-Skyrms model is given in Fig. 6.

This curve is a very simple object: looking at the point the curve touches the y-axis, we can tell that if we wish to reach an expected distortion of 0 we need a rate that matches the entropy of states, i.e., $\log_2 3 \approx 1.58$ bits. This is because the only way for the encoder-decoder pair to get it always right is to have a *signaling system* (Huttegger 2007, proposition 3) in which the sender chooses a

[6] The expected distortion D is the average of distortion values for each pair of a state S_i and an act A_j, d_{ij}, weighted by the joint probability of those state and act, $\Pr(S_i, A_j)$.

[7] This way of calling it comes from Bruner et al. (2018).

	A_1	A_2	A_3
S_1	1	0	0
S_2	0	1	0
S_3	0	0	1

Table 1: Payoff matrix for the 3x3 Lewis signaling game

	A_1	A_2	A_3
S_1	0	1	1
S_2	1	0	1
S_3	1	1	0

Table 2: Distortion measure for the 3x3 Lewis signaling game

different signal for each state and the receiver chooses the right act in face of each signal. The entropy of signals (thus, the rate of the code) in this situation is the aforementioned 1.58 bits.

In fact one can directly argue for that rate value at distortion 0 on purely information-theoretic grounds: Shannon's *source coding theorem* (Cover & Thomas 2006, Theorem 7.13.1) entails that there exists a channel which has a rate equal to or larger than the entropy of the source and an error probability arbitrarily close to zero, and that no code with lower rate can achieve this. Signaling systems are a consequence of the source coding theorem.

On the other hand, if the rate is zero (that is, if the channel is completely closed, and sender and receiver do not communicate) the best achievable expected distortion is $2/3 \approx 0.66$. This is achieved, e.g., by the receiver always doing A_1, no matter what. This act will achieve a distortion of 0 one third of the time (whenever S_1 is the case, remember that all three states are equiprobable) and a distortion of 1 two thirds of the time, which adds up to an expected distortion of 0.66. Distortions in between 0 and 1 correspond to different rates, as the curve shows.

The rate-distortion function, as depicted in Fig. 6, fully characterizes the informational structure of the 3x3 Lewis signaling game. Interrogating it is useful. First, is there a problem that transmitting information can help solve? Yes, there is: there are levels of distortion (indeed, distortion zero) which are achievable with information transfer (i.e., with codes of nonzero rate, and therefore channels of nonzero capacity) and not achievable otherwise. Second, is there a distortion (equivalently, payoff) optimum for both sender and receiver? Yes, there is as well: that would be the point at the upper left corner of the plot, where distortion zero is achieved with a 1.58 bit code.

The rate-distortion curve is all there is to the informational problem at hand. The informational facts represented in the curve, unsurprisingly, explain a lot of the sender-receiver goings-on in the 3x3 Lewis signaling game: for example, the

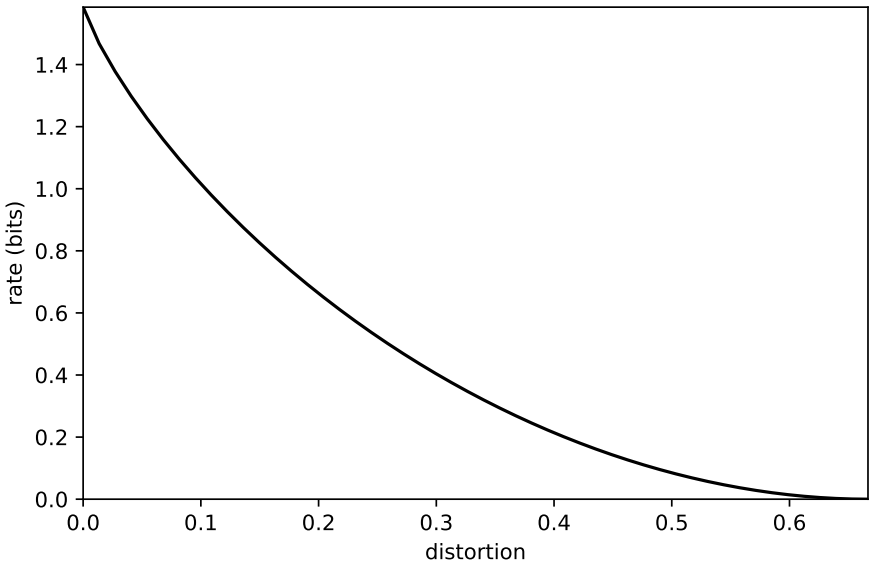

Figure 6: The Rate-Distortion Curve for the 3x3 Lewis Signaling Game

only strict Nash equilibria and the only evolutionarily stable strategies in $n \times n$ Lewis signaling games are distortion zero / maximum rate points. This follows from all such equilibria being signaling systems (Huttegger 2007, propositions 4 and 5) and all signaling systems constituting zero-distortion codes.[8]

One natural (but, as far as I am aware, unexplored) way to think of signaling games is as, first and foremost, *rate-distortion problems faced by sender and receiver*, where Nash equilibria and dynamical sinks describe ways to approximate optimal solutions to these problems without joint deliberation. Game-theoretic discussion of signaling is often cast in confrontational terms. The rate-distortion perspective allows us to see sender-receiver behavior as much more cooperative than it is typically seen.

So far I have only dealt with a very simple case, in which sender and receiver have perfect common interest. I will now substantiate the foregoing remarks by showing how the rate-distortion perspective can be extended to cases of imperfect common interest.

[8]On the other hand, certainly not everything is explained by the rate-distortion curve: in the same remarkable paper, Simon Huttegger also proves that not all random starting points subjected to an evolutionary regime governed by the two-population replicator dynamical equations reach a zero-distortion / maximum rate code—although only such codes correspond to aymptotically stable points. In the 3x3 case, the proportion that don't reach distortion zero amount to just below 5% in numerical trials (Huttegger et al. 2010). See Hofbauer & Sigmund (1998) for a description of the replicator dynamics, and Martínez & Godfrey-Smith (2016) for a gentler introduction.

4 Imperfect common interest

It is possible to present the 3x3 Lewis signaling game as a curve, as in Fig. 6, because there is only one distortion objective: that is, both sender and receiver agree completely in the distortion measure. Still, if one so wishes, the same information can be presented by making it explicit that both sender and receiver have their own distortion measure—it's just that both coincide. One way to do it is with a two-dimensional heatmap, as in Fig. 7.

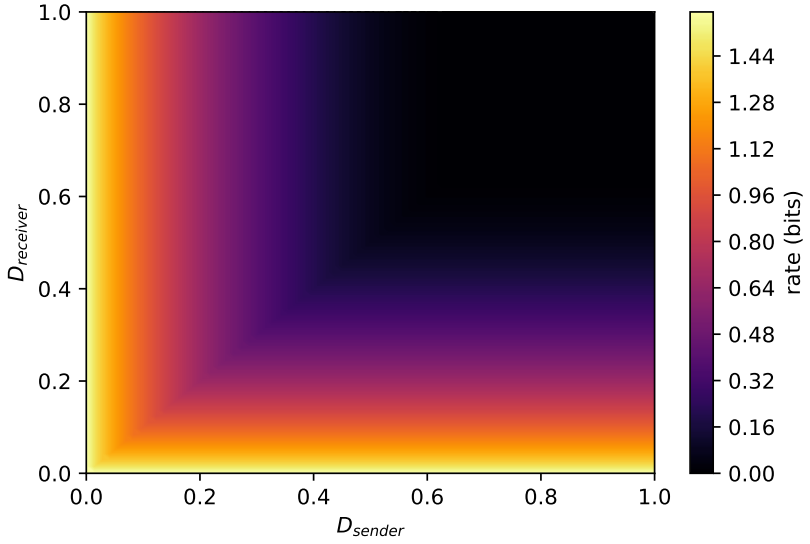

Figure 7: Rate-Distortion surface for a 3x3 Lewis signaling game

The way to read this plot is as follows: the x-axis codes distortion goals for the sender, and the y-axis, for the receiver. The color at each point in the plot codes the rate necessary to meet, or exceed, both distortion goals. So, for example, the black area in the upper-right area of the plot delimits all pairs of sender- and receiver-distortion goals that can be met with zero rate—i.e., with no information at all passing from sender to receiver. As we saw above, any point with D_{sender} and $D_{receiver} \leq 2/3$ can be reached with no information transfer whatsoever. Below and to the left of the black area, the lower the distortion goals are, the higher the rate necessary to achieve them. As we saw above, it is possible to achieve zero expected distortion for both sender and receiver (that would be the lower left corner), and the rate necessary to achieve that point is $\log_2 3 \approx 1.58$ bits—the pale-yellow end of the color-coded rates.[9] The fact that, in Fig. 6, the

[9] Why does it take 1.58 bits to reach the upper left and lower right corners? Because the only way to achieve these goals is by achieving zero distortion for both players. Remember that, as per eq. 4, distortion goals have to be at least met, but can be exceeded.

rate-distortion curves becomes increasingly steeper as we approach distortion zero is mirrored in Fig. 7 by colors 'heating up' faster near the lower-left corner than near the upper-right one.

Fig. 7 is representing, in a less economical way, the situation already represented in fig. 6. A two-dimensional heatmap is unnecessary when sender and receiver payoff matrices (distortion measures) coincide. Indeed, this is the scenario typically studied in information theory: one in which encoder and decoder cooperate to achieve a common communication goal. The rate-distortion approach could hardly be recommended as an addition to the signaling-game theorist's toolbox if it only worked for cases of perfect common interest. Fortunately, the extension of rate-distortion analyses to cases of imperfect common interest is entirely conservative and straightforward, if seldom explored. In his seminal paper on the numerical computation of rate-distortion functions, Richard Blahut discusses the situation in which "it may be desired that two (or more) separate definitions of distortion be satisfied" (1972, p. 470). The use case he mentions is when "the reproduced data is to be made available to two different users with different applications in mind." (*ibid.* p. 471). To the best of my knowledge, exploration of this proposed extension of rate-distortion theory to two distortion measures is virtually non-existent in the information-theoretic literature. In any case, I believe the present article to be the first to explore it in the context of Lewis-Skyrms signaling games.

The way to adapt Blahut's idea to signaling games is, quite simply, to convert the two payoff matrices for sender and receiver into two independent distortion measures, d_s for the sender and d_r for the receiver, and consider the case in which a distortion objective for the sender, D_S, and another for the receiver, D_R, must be satisfied (or exceeded) jointly. When there is divergence of interest, then, the rate-distortion curve becomes a rate-distortion surface: we define

$$\delta_s(s,a) = \sum_{(s,a)} \Pr(s,a) d_s(s,a) \tag{5}$$

and

$$\delta_r(s,a) = \sum_{(s,a)} \Pr(s,a) d_r(s,a) \tag{6}$$

The rate-distortion function for two independent distortion measures then becomes

$$R(D_S, D_R) = \min_{\{\Pr(a|s):\delta_s(s,a)\leq D_S, \delta_r(s,a)\leq D_R\}} I(S;A) \tag{7}$$

The rate-distortion surface presents the informational problem that sender and receiver must solve: minimizing distortion simultaneously for two users with

different needs. Take, for example, the main game Brian Skyrms uses to illustrate his account of deception (Skyrms 2010, p. 81). The payoff matrices for sender and receiver are given in Table 3. Each payoff matrix corresponds, in the manner described above, to a distortion measure. The two distortion measures and the fact that the source consists in three equiprobable states (*ibid.*) leave us with the rate-distortion surface in Fig. 8. The bottom-left white region is *unreachable*: there is no code that can transmit information in a way such that both D_S and D_R can be met, for points $\langle D_S, D_R \rangle$ in that region. Every other point is *reachable*.[10]

	A_1	A_2	A_3
S_1	2, 10	0, 0	10, 8
S_2	0, 0	2, 10	10, 8
S_3	0, 0	10, 10	0, 0

Table 3: Payoff matrices for sender and receiver in Skyrms's deception game.

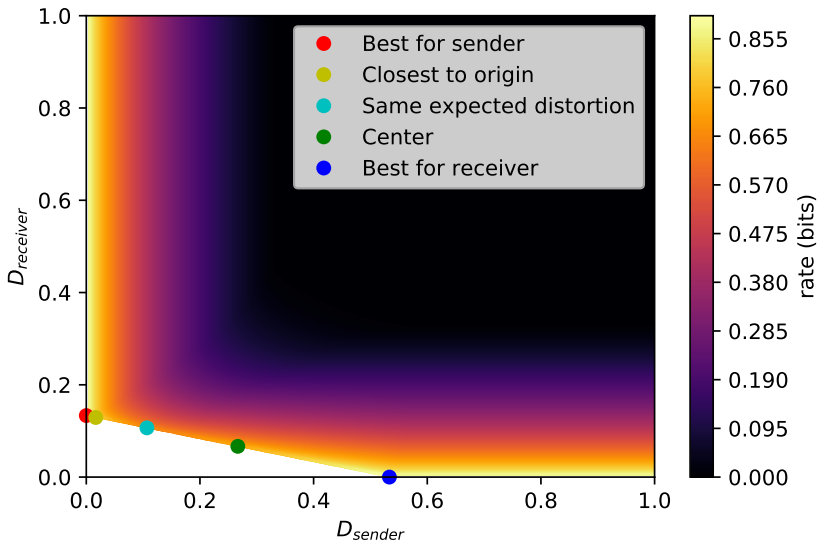

Figure 8: Rate-Distortion surface for Skyrms's deception game. The colored dots are notable points in the Pareto frontier, characterized in the main text.

[10] I state, without proof, a fact about rate-distortion surfaces that is useful in constructing them. The proof, which is easy if somewhat involved, will be provided in subsequent work.

Let us say that a point $\langle D_S, D_R \rangle$ is *strictly reachable* iff there is a code with exactly those expected distortions. I will also call codes $\Pr(a|s)$ for which all $\Pr(a_j|s_i)$ are either 0 or 1 *deterministic*. It can then be shown that

A point $\langle D_S, D_R \rangle$ is strictly reachable iff it corresponds to the expected distortions of a convex combination of deterministic codes.

It is easy to see that every point that is reachable tout court is reachable by a sender and a receiver playing a signaling game—just not necessarily reachable with the minimum achievable rate. Suppose that $\langle D_S, D_R \rangle$ is reachable. Then there is a Lewis-Skyrms code C that has these, or lower, distortion values. We just need to find a sender strategy, S, and a receiver strategy, R, such that $S \cdot R = C$. There are two trivial solutions to this equation: one is to set $S = I_s$ and $R = C$, where I_s is the $s \times s$ identity matrix; the other is to set $S = C$ and $R = I_a$, where I_a is the $a \times a$ identity matrix.

Again here, we can interrogate the rate-distortion surface to learn about the underlying informational problem: is this a game in which both sender and receiver stand to gain by opening the channel? Yes: there are points to the left and below the zero-rate black region. That is, points reachable by a code of nonzero rate, where both sender and receiver have less expected distortion than anything they can achieve in the zero-rate region.

The best reachable points for both sender and receiver lie in the straight line connecting $\langle 0, 0.133 \rangle$ (the red dot in Fig. 8) with $\langle 0.533, 0 \rangle$ (the blue dot in Fig. 8). That is, the frontier between reachable and unreachable. Any point to the right of or above this line has a counterpart on the line that has a distortion at least not worse for both players and strictly better for at least one. A set of points meeting this condition is sometimes called a *Pareto frontier*. I will follow this usage here.

Now, *which* point precisely in the Pareto frontier is the fairest solution to the problem of, paraphrasing Blahut, "minimizing losses to two different users with different applications in mind", is undefined. A few plausible candidates are:

Center: The center of the Pareto frontier $\langle 0.267, 0.067 \rangle$

Same Expected Distortion: The point in the Pareto frontier at which the expected distortion for both players is the same $\langle 0.107, 0.107 \rangle$.

Closest to Origin: The point in the Pareto frontier that is closest to the unreachable no-distortion point $\langle 0, 0 \rangle$, i.e., $\langle 0.017, 0.130 \rangle$.

Any of these points, and probably a handful more, could plausibly be chosen by a fully cooperative team in charge of designing a code that accommodates the needs of sender and receiver.[11]

Skyrms (2010, chap. 6) offers an account of deception that falls in the family of what Fallis and Lewis call *functional deception*: "the view that deception only requires that a deceiver *benefit* from sending a *misleading* signal" (Fallis & Lewis 2017, p. 3). In Skyrms's own development of this idea, first, a signal M_j is *misleading* iff there is a state S_i such that either the world is not in S_i

[11] The rate at these points (that is, how wide the channel needs to be at the best solution to the bargaining problem sender and receiver face) could plausibly be regarded as a measure of common interest between them. Comparing it to the measures of common interest described in (Godfrey-Smith & Martínez 2013; Martínez & Godfrey-Smith 2016) is matter for another paper.

and the information M_j carries about S_i is positive, or the world is in S_i and the information M_j carries about S_i is negative. For the notion of carrying information about a state see formula (1) above. Second, the sender benefits from sending this misleading signal if "they do better than they would have had the receiver known the true state with probability 1" (paraphrased from Fallis & Lewis 2016, p. 8)

In his discussion of deception Brian Skyrms focuses on a Nash equilibrium in the signaling game in Table 3 that results in the code marked with a red dot in Fig. 8:

> **Best for Sender:** If sender send signal M_1 in states S_1 and S_2 and signal M_2 in state S_3; if receiver do act A_3 on receipt of signal M_1 and act A_2 on receipt of signal M_2. (Skyrms 2010, p. 81—variables changed).

In the game of Table 3 and this Best for Sender equilibrium above, signal M_1 is deceptive in Skyrms's sense. It misleads: when sent in S_1 (S_2) it raises the probability of S_2 (S_1). As a consequence, it carries positive information about a non-actual state. This is to the detriment of the receiver, who is forced to do A_3, the best cover-all act for the receiver for S_1 and S_2, but not the best act for the receiver for either S_1 or S_2 individually.

I do not wish to contest Skyrms' definition of deception. The functional-deception tradition is surely right that misleadingness plus benefit captures an important part of what we mean by deception. But I do wish to contest that a confrontational description of what happens in this game, for example, one in terms of manipulation, is the most apt one. One can perfectly see "deceptive" results as emerging from a cooperative endeavor—hence the somewhat provocative title of this piece. In particular, it turns out that the candidates for a fair compromise in respecting sender and receiver interests identified above (Center, Same Expected Distortion, and Closest to Origin) come out as deceptive and manipulative, according to Skyrms's treatment. I draw this out in the following section.

5 Deception as Cooperation

To recap, the cooperative endeavor I was referring to above is that of *constructing a code that simultaneously minimizes sender and receiver distortion measures.* A sender strategy is just an encoder; a receiver strategy, just a decoder. The most important behavioral unit is neither of those strategies, but the resulting encoding-decoding pair—the code that connects states to acts. By the same token, it is not a good idea to base confrontational and manipulative descriptions upon the behavior of individual signals, because all signals conspire to generate a code, and it is full codes that senders and receivers care about. Individual signals are just means to an end.

For example, it turns out that the Best for Sender equilibrium above represents a very good solution, for both parties, to this cooperation problem. It is in

the Pareto frontier, and very close to the Closest to Origin point—the receiver only does 2.5% worse at this point than at Closest to Origin. Moreover, at this point, the rate of the code is $H(1/3) = 0.92$. Rate, recall eq. 6, corresponds to the minimal mutual information between states and acts, $I(S;A)$, at which the two distortion goals are jointly achievable.[12] 0.92 bits is the maximum information transfer anywhere in the R-D surface. That is, no reachable point needs a higher mutual information between states and acts; and, in particular, no more information is needed for the receiver to achieve zero distortion. The problem with Best for Sender, to the extent that there is one, does not seem to be a lack of informativeness on the part of the sender.

In any event, my case does not much hang on Best for Sender not being particularly manipulative, because *the center of the Pareto frontier is also Skyrms-deceptive.* Figure 9 gives a code that sits at exactly the Center point (the numbers in the matrix are conditional probabilities of acts on states, as per Fig. 5):

$$\begin{bmatrix} .3 & 0 & .7 \\ 0 & .7 & .3 \\ 0 & 1 & 0 \end{bmatrix}$$

Figure 9: The code corresponding to the Center point

As one can see by comparing this code to the payoff matrix in Table 3, the usefulness of this code for sender and receiver is exquisitely balanced, so that both get exactly the same payback out of setting it up. Yet, there are straightforward ways of implementing this code with Skyrms-deceptive signals. One such way is one of the two trivial encoding-decoding pairs for any given code, described above—the one in which the receiver strategy is the identity matrix:

$$\begin{bmatrix} .3 & 0 & .7 \\ 0 & .7 & .3 \\ 0 & 1 & 0 \end{bmatrix} \qquad\qquad \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Figure 10: Sender and receiver strategies that implement the code in Fig. 9

Here, signals M_2 and M_3 are Skyrms-deceptive. Take M_3. The probabilities of states conditional on M_3 are $[.7, .3, 0]$. Remember that we are dealing with equiprobable states, so the unconditional probabilities of states are $[1/3, 1/3, 1/3]$. M_3 is therefore misleading when sent in S_2, according to Skyrms's understanding of misleadingness: $\Pr(S_2|M_3) = 0.3 < P(S_2) = 0.33$. This misleadingness is also to the benefit of the sender: the receiver performs A_3, the best cover-all act for states S_1 and S_2, instead of A_2, the best act for the receiver in S_2.[13]

[12] In Godfrey-Smith & Martínez (2013, p. 3) we offer a justification, aimed at convincing game-theorists, for using $I(S;A)$ as the measure of meaningful information transfer. This is anyway, as we have seen, the standard in information theory.

[13] Fallis and Lewis (2016, 2017) have produced convincing counterexamples to Skyrms's

It is somewhat awkward to claim that the sender is manipulating the receiver in this situation, when the net result of their behavior has been *designed* to be maximally fair—and if you are not convinced that Center is maximally fair, the same argument could be repeated for other points in the Pareto frontier. Whatever manipulation happens from sender to receiver must be offset by the exact same amount of manipulation from receiver to sender.

Deception, in the functional-deception literature, is entirely a sender-to-receiver affair. Receivers cannot deceive senders. This is reasonable: messages go from sender to receiver, and it's hard to see how the receiver could deceive in retrospect. But manipulation is another thing entirely. The receiver *does* have the resources to manipulate the sender, by granting or withdrawing access to certain actions. For example, the blue dot in Fig. 8, at which the receiver has zero distortion, can also be reached with a Nash equilibrium:

Best for Receiver: If sender send signal M_1 in state S_1 and signal M_2 in states S_2 and S_3; if receiver do act A_1 on receipt of signal M_1 and act A_2 on receipt of signal M_2.

This equilibrium is the mirror image of the Best for Sender equilibrium. The informational properties of the two equilibria are entirely analogous. The only difference is that in the former it's the sender that stands to gain; in the latter, the receiver. It is perhaps awkward to claim that the receiver is deceiving the sender here, but, regarding manipulation, there is no such awkwardness: either the receiver is manipulating the sender in the latter equilibrium, or no one is manipulating anyone in neither equilibrium.

---

analysis of misleadingness, and suggest that other measures of epistemic utility present in the literature, such as the *Brier rule*, the *logarithmic* rule or the *spherical* rule, should be examined and might be preferable (see Fallis & Lewis 2016, p. 579 for details) Their idea is to deem a signal, M_i, misleading iff the epistemic utility of the probability distribution over states conditional on M_i is lower than the epistemic utility of the unconditional probability distribution over states (Fallis & Lewis 2017).

All three of these epistemic-utility rules agree that, in the case I have been discussing, the vector of state probabilities conditional on M_3 has lower epistemic utility than the vector of unconditional probabilities. Using Δ to refer to the difference between the epistemic utility of state probabilities conditional on M_3 and the utility of unconditional probabilities, if S_2 is the actual state (so that $\Delta < 0$ corresponds to a misleading message), $\Delta_{Brier} = -.313$, $\Delta_{logarithmic} = -.106$, $\Delta_{spherical} = -.183$. M_3 comes out misleading also according to Fallis and Lewis's criterion.

One final point regarding the misleadingness of signals: in both Skyrms's and Fallis and Lewis's analyses misleadingness depends only on the joint probability of states and signals. Payoffs are ignored. Yet, it is reasonable to think that being presented with a distorted image of the *payoff* structure of the world is more important to the receiver than being presented with a distorted image of its *probabilistic* structure—although the two are obviously related. An investigation of how epistemic utility relates to, well, utility is also matter for another paper.

6 Conclusions

Both *Best for...* Nash equilibria are reasonable compromises for both players. Both lie in the Pareto frontier, and no other point in the frontier, *a fortiori* none of the fairer options discussed above, is reachable by a Nash equilibrium. In any event, functional-deception analyses are unable to distinguish these points from exquisitely egalitarian, non exploitative strategies of information transmission with two different distortion measures, such as Center.

Some of the confrontational rethoric that typically goes with analyses of deception is perhaps a nod to the 'manipulationist' approach to communication spearheaded by Dawkins and Krebs (1978; Krebs & Dawkins 1984; see also Adams & Caldwell 1990; Byrne & Whiten 1990; Endler 1993; Owings & Morton 1997), according to which communication is just "a means by which one animal makes use of another animal's muscle power" (Dawkins & Krebs 1978, p. 283). Whatever the merits of this approach, manipulativeness cannot be established solely on the basis of the behavior of individual signals. Signals are just a means to the end of building a *code* that translates information about states into acts. An individual signal only makes sense in the context of its code. Furthermore, information is not a neutral commodity. Getting some things right is more important than getting others right, differently so for each of the interested parties. Factoring these observations in is, I submit, central to understanding the informational structure of signaling games. Here I have offered rate-distortion analyses as an obvious way to do so.

Funding

This work is supported by the Spanish Ministerio de Economía, Industria y Competitividad, through grant RYC-2016-20642, and by the Generalitat de Catalunya, through grant 2017-SGR-63.

References

Adams, ES & Caldwell, RL 1990, 'Deceptive communication in asymmetric fights of the stomatopod crustacean Gonodactylus bredini', *Animal Behaviour*, vol. 39, no. 4, pp. 706–716.

Arimoto, S 1972, 'An algorithm for computing the capacity of arbitrary discrete memoryless channels', *IEEE Transactions on Information Theory*, vol. 18, no. 1, pp. 14–20.

Blahut, R 1972, 'Computation of channel capacity and rate-distortion functions', *IEEE transactions on Information Theory*, vol. 18, no. 4, pp. 460–473.

Bouma, G 2009, 'Normalized (pointwise) mutual information in collocation

extraction', *Proceedings of GSCL*, pp. 31–40.

Bruner, J, O'Connor, C, Rubin, H & Huttegger, SM 2018, 'David Lewis in the lab: Experimental results on the emergence of meaning', *Synthese*, vol. 195, no. 2, pp. 603–621.

Byrne, RW & Whiten, A 1990, 'Machiavellian Intelligence: Social Expertise and the Evolution of Intellect in Monkeys, Apes, and Humans', *Behavior and Philosophy*, vol. 18, no. 1, pp. 73–75.

Cover, TM & Thomas, JA 2006, *Elements of Information Theory*, New York: Wiley.

Dawkins, R & Krebs, JR 1978, 'Animal signals: Information or manipulation', *Behavioural ecology: An evolutionary approach*, vol. 2, pp. 282–309.

Dretske, F 1981, *Knowledge and the Flow of Information*, The MIT Press.

Endler, JA 1993, 'Some general comments on the evolution and design of animal communication systems', *Phil. Trans. R. Soc. Lond. B*, vol. 340, no. 1292, pp. 215–225.

Fallis, D & Lewis, PJ 2016, 'The Brier rule is not a good measure of epistemic utility (and other useful facts about epistemic betterness)', *Australasian Journal of Philosophy*, vol. 94, no. 3, pp. 576–590.

Fallis, D & Lewis, PJ 2017, 'Toward a formal analysis of deceptive signaling', *Synthese*, pp. 1–25.

Godfrey-Smith, P & Martínez, M 2013, 'Communication and Common Interest', *PLOS Computational Biology*, vol. 9, no. 11.

Hofbauer, J & Sigmund, K 1998, *Evolutionary Games and Population Dynamics*, Cambridge: Cambridge University Press.

Huttegger, SM 2007, 'Evolution and the Explanation of Meaning', *Philosophy of Science*, vol. 74, no. 1, pp. 1–27.

Huttegger, SM, Skyrms, B, Smead, R & Zollman, K 2010, 'Evolutionary Dynamics of Lewis Signaling Games: Signaling Systems vs. Partial Pooling', *Synthese*, vol. 172, pp. 177–191.

Kraemer, DM 2015, 'Natural probabilistic information', *Synthese*, vol. 192, no. 9, pp. 2901–2919.

Krebs, JR & Dawkins, R 1984, 'Animal signals: Mind-reading and manipulation', *Behavioural ecology: an evolutionary approach, 2nd edn (ed. JR Krebs & NB Davies)*, pp. 380–402.

Lewis, D 1969, *Convention: A philosophical study*, John Wiley & Sons.

MacKay, DJ 2003, *Information theory, inference and learning algorithms*, Cam-

bridge university press.

Martínez, M 2015, 'Deception in SenderReceiver Games', *Erkenntnis*, vol. 80, no. 1, pp. 215–227.

Martínez, M & Godfrey-Smith, P 2016, 'Common Interest and Signaling Games: A Dynamic Analysis', *Philosophy of Science*, vol. 83, no. 3, pp. 371–392.

Owings, DH & Morton, ES 1997, 'The role of information in communication: An assessment/management approach', *Communication*, Springer, pp. 359–390.

Scarantino, A 2015, 'Information as a probabilistic difference maker', *Australasian Journal of Philosophy*, vol. 93, no. 3, pp. 419–443.

Shannon, C 1948, 'A Mathematical Theory of Communication', *The Bell System Mathematical Journal*, vol. 27, pp. 379–423, 623–656.

Shannon, CE 1959, 'Coding theorems for a discrete source with a fidelity criterion', *IRE Nat. Conv. Rec*, vol. 4, no. 142-163, p. 1.

Shannon, CE & Weaver, W 1998, *The Mathematical Theory of Communication*, University of Illinois press.

Shea, N 2007, 'Consumers Need Information: Supplementing Teleosemantics with an Input Condition', *Philosophy and Phenomenological Research*, vol. 75, no. 2, pp. 404–435.

Skyrms, B 2010, *Signals: Evolution, Learning & Information*, New York: Oxford University Press.

Stegmann, U 2015, 'Prospects for Probabilistic Theories of Natural Information', *Erkenntnis*, vol. 80, no. 4, pp. 869–893.