UNIVERSITY OF CALIFORNIA,
IRVINE

Deception and Formal Models of Communication

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Philosophy

by

Gregory McWhirter

Dissertation Committee:
Distinguished Professor Brian Skyrms, Chair
Professor Simon Huttegger
Professor Jeffrey A. Barrett

2014

DEDICATION

To my parents for their continued support and encouragement.

Contents

List of Figures

List of Tables

ACKNOWLEDGMENTS

Some material is based upon work supported by the National Science Foundation under Grant No. EF 1038456 administered by Simon Huttegger. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

I would like to thank, in particular, Rory Smead, Brian Skyrms, Simon Huttegger, Jeffrey Barrett, Justin Bruner, Cailin O'Connor, Hannah Rubin, and several anonymous referees from the *British Journal for the Philosophy of Science* for their helpful discussions and for providing many useful comments and questions in preparing and revising this dissertation.

CURRICULUM VITAE

Gregory McWhirter

EDUCATION

Doctor of Philosophy in Philosophy **2014**
University of California, Irvine *Irvine, California*

Master of Arts in Mathematical Behavioral Science **2011**
University of California, Irvine *Irvine, California*

Bachelor of Science in Mathematics-Philosophy **2008**
University of Pittsburgh *Pittsburgh, Pennsylvania*

SELECTED HONORS AND AWARDS

Graduate Research Fellowship Honorable Mention **2009**
National Science Foundation

ABSTRACT OF THE DISSERTATION

Deception and Formal Models of Communication

By

Gregory McWhirter

Doctor of Philosophy in Philosophy

University of California, Irvine, 2014

Distinguished Professor Brian Skyrms, Chair

Having a satisfactory definition of behavioral deception is important for understanding several types of evolutionary questions. No definition offered in the literature so far is adequate on all fronts. After identifying characteristics that are important for a definition, a new definition of behavioral deception is offered. The new definition, like some other proposed attempts, relies on formal game-theoretic models of signalling. Unlike others, it incorporates explicit consideration of the population in which the potentially deceptive interactions occur. The general structure of the definition satisfies many of the characteristics that were problematic for other definitions, and others are satisfied by explicitly incorporating information about the population in which interactions occur.

The proposed definition is applied to chick begging in the Sir Philip Sidney game and stomatopod bluffing behavior. The definition is shown to allow universal deception in equilibrium, contrary to claims by Kant that such a thing should be impossible. An extension is also considered to investigate the potential evolutionary advantages of self-deception.

Chapter 1

Introduction

Attempts to define deception have proceeded along two distinct paths. On the one hand, there are traditional philosophical attempts, focusing on beliefs and intentions of the agents involved. This has sometimes been called an intentional approach. On the other hand, there are behavioral approaches. These attempts try to focus exclusively on observed behavior to identify deception, disregarding any beliefs or intentions that might be present.

The most prominent definitions of behavioral deception are all in some way inadequate.[1] I will suggest, instead, that formal modelling of behavior should be used for the identification of behavioral deception. Several attempts in this vein have already been made by Lachmann and Bergstrom (2004) and Skyrms (2010), among others. Although these attempts employ formal models, each has its own flaws.

[1] Each of them is possibly applicable in certain circumstances that are favorable. My argument will be that they are *in general* problematic to apply.

1.1 Deceptive Animal Behavior

Many examples of behavior that we might want to describe as deceptive have been described by a variety of authors and in a variety of species.

Amotz Zahavi, in the context of mate choice, suggests that systematic deception in mate selection signalling is prevented through a system of handicaps. (Zahavi, 1975, 1977) Zahavi argues that employing signals tied to male quality prevents males from deceptively indicating that they are of a superior quality than they really are. Any individuals trying to indicate that they are of a higher type than they are would enjoy lower fitness due to the increased signal costs.

Dorothy Cheney and Robert Seyfarth provide two examples of behavior that we would, perhaps even more obviously, want to call deceptive. One of these was observed in vervet monkeys and the other in baboons. In an observational study of vervet monkeys, Cheney and Seyfarth noted several occasions when Kitui, a low ranking male, falsely gave an alarm call, seemingly to prevent the transfer of another male into the troop. (Cheney and Seyfarth, 1992, pp. 213–4) In another study of baboons, they describe the following situation:

> Hannah, the seventh-ranking female at the time, had been receiving attention all morning from higher-ranking females who wanted to handle her baby. Although these females had always been friendly, their constant attentions had prevented Hannah from eating or resting. Hannah had just sat down to eat a fig when Sierra, the third-ranking female, approached and reached for her baby. Hannah grabbed Sierra's hand and cuffed her on the face. Although Hannah's threat violated the established rank order, Sierra did not retaliate but moved away. An hour later, Sierra approached Hannah again. Perhaps remembering that she had hit Sierra earlier, Hannah flinched and began to move away, but she relaxed

> when Sierra began to grunt [(a typical sign of nonaggression)]. As soon as Sierra reached Hannah, she lept on her and bit her on the neck. (Cheney and Seyfarth, 2008, p. 154)

In both of these examples, we would be quite tempted to say that one of the vervets or baboons was being deceptive. Kitui and Sierra appeared to be giving signals appropriate for certain situations when in fact those situations were not the case.

There are many more examples of animal behavior we might want to call deceptive, from bluffing in stomatopod confrontations (Steger and Caldwell, 1983), to false alarm calls in mixed-species bird flocks (Searcy and Nowicki, 2005, p. 65) and Batesian mimicry in a variety of species (see, e.g., Bond and Robinson, 1988).

Having a standard definition of deception that is applicable to the vast majority of cases like these is desirable for several reasons. First, having a common definition should reduce misunderstandings based on varying intuitions or underspecified anthropomorphic analogies. For example, one might interpret the situation involving Kitui described briefly above in two conflicting ways. On the one hand, Kitui could be interpreted as deceptively giving a false alarm call because the signal for a predator is given in the absence of that predator. On the other hand, he could be interpreted as not being deceptive if one understands the vocalization as indicating that the sender wants the receiver to run up a tree and out onto the end of a branch.[2] Second, having a widely applicable definition of deception can aid in an effective investigation of some interesting evolutionary questions. One example of such a question, taken up by Trivers and others, is whether self-deception is evolutionarily advantageous. (see, e.g., Trivers, 2000; von Hippel and Trivers, 2011)

Standard philosophical definitions of deception in human communication involving belief,

[2]The particular action of running up a tree and out onto a branch would be the optimal sort of response if there was a leopard present, as the leopard could not follow them. This action would also prevent the incoming male from integrating with the new troop, as that male would respond by running into a tree near the troop he was leaving.

intent, etc. are inapt for application to these scenarios. Moreover, even if it were possible to apply them, they are not generally necessary to address many interesting questions. Searcy and Nowicki explain their preference – with which I agree – for behavioral definitions through their relevance in evolutionary contexts and through the irrelevance of mental states.

> [W]e are interested in how natural selection shapes animal communication to be either honest or dishonest. From this viewpoint, the question of mental states is largely irrelevant; the costs and benefits to the signaler of giving a false alarm, and to the receiver of responding, ought to be the same whether or not the signaler is able to form an intention and the receiver to form a belief. (Searcy and Nowicki, 2005, p. 5)

1.2 Behavioral Definitions of Deception

Several attempts have been made to provide this sort of definition relying only on behavioral descriptions. An early synthesis from strains of sociobiological thought was given by Robert Mitchell:

Definition 1.1 (Mitchell)**.** [Deception occurs when]

(i) An organism R registers something Y from organism S;
(iia) R acts appropriately toward Y, because
(iib) Y means X; and
(iii) it is untrue that X is the case.

(Mitchell, 1986, p. 20)

This definition is not acceptable for two reasons. The first problem is one that Mitchell himself recognized: this definition does not distinguish between deception and mere error on

the part of the organism providing the signal Y.[3] The second problem is more severe, but only appears in the context of Mitchell's presentation. In this definition, the something Y that is registered is supposed to be something actively provided: "deception involves providing rather than retaining information." (Mitchell, 1986, p. 17) One need not limit the definition in this way, and in fact, it is probably better not to. There are many situations we would want to call deceptive that involve *not* providing information, such as a chimpanzee not signalling that he or she had found a source of food.

John Maynard Smith and David Harper offer a different definition:

Definition 1.2 (Maynard Smith and Harper)**.** Consider a signal that is given in more than one circumstance, but always produces the same response in receivers. Receivers usually benefit from their response, but deception can occur if there is another circumstance in which the same response benefits the signaller at the receiver's expense. (Maynard Smith and Harper, 2003)

Maynard Smith and Harper attribute this definition to Stuart Semple and Karen McComb. Although the exact wording is different, the sentiment is the same:

Definition 1.3 (Semple and McComb)**.** An interaction qualifies as behavioural deception when, as a result of the behaviour of the signaller, the receiver registers a certain situation that is not in reality occurring. As a result of the interaction, the signaller benefits, while the receiver pays a cost. (Semple and McComb, 1996, p. 434)

William Searcy and Stephen Nowicki offer their own, slightly different, but largely similar definition:

Definition 1.4 (Searcy and Nowicki)**.** ... we will define deception as occurring when:

[3]Mitchell gives a revised definition that includes signaller benefit. He alters requirement (i) to read "An organism R registers (or believes) something Y from some organism S, where S can be described as benefiting when (or desiring that)." This revised definition is similar to those that follow, with some minor variations in wording.

1. A receiver registers something Y from a signaller;
2. The receiver responds in a way that
   (a) benefits the signaler and
   (b) is appropriate if Y means X; and
3. It is not true here that X is the case.

(Searcy and Nowicki, 2005, p. 5)

All three of these definitions successfully differentiate deception from mere error.[4] The main difference is whether detriment to the receiver is required for there to be deception. Skyrms does not think this distinction is important. (see Skyrms, 2010, 75–6 fn. 5) The actual problems with these definitions do not hang on this possible distinction, though.

Although these definitions are largely effective, each also includes some components that are insufficiently clear. In particular, they reference 'meaning' or regularities that might be difficult to interpret in a given situation. In the example of Kitui, does the "false" alarm call mean "there is a leopard", or does it mean "I think you should run up a tree"? If it means the former, then the situation could very well be deceptive according to Searcy and Nowicki's definition; if it means the latter, then it is probably not deceptive.[5]

That particular example isn't obviously a problem for Maynard Smith and Harper's definition. Other parts of their definition, however, are similarly unclear or problematic. Their definition, for example, includes the condition "at the receiver's expense". It is unclear what counts as an expense in this sense. It is contrasted with receiving a benefit, but does receiving a lesser benefit than one might have otherwise received qualify as an expense, or must

[4]These definitions cannot separate deception from 'lucky' errors, only from non-beneficial errors. The definition I will later propose does not differentiate these either. This is not a problem, in my opinion. I will be content to identify 'lucky' errors as deceptive.

[5]In fact, I think attributing propositional content to such signals is probably a mistake altogether. (see also Skyrms, 2010, pp. 40–4) This does not remove the confusion in the given definition, however. There are several other positions on this topic as well. (see, e.g., Harms, 2004; Millikan, 2004)

the expense be a loss below some fixed baseline?

Although it is not guaranteed that no informal definition can be completely acceptable, none of those offered so far have been. Switching contexts to formal models will allow a clearer and more widely applicable definition of deception to emerge.

1.3 Formal Models of Signalling

The definition of deception I propose and the applications I will examine for it rely on formal models of signalling. The formal models commonly used in modelling animal communication are modifications of the signalling games described by David Lewis (Lewis, 1969). These are extensive-form games of imperfect information.

In the simplest case, there are two players: the sender and the receiver. Initially, Nature probabilistically chooses a state of the world.[6] The sender observes the state of the world and sends a message to the receiver. The possible messages are defined beforehand, but have no prior meaning; they can be thought of as arbitrary but distinguishable noises, flashes of color, or motions. The receiver observes the message that was sent, but does not observe the state of the world. Based on the observation of the message, the receiver chooses an action to perform from a set of possible actions defined beforehand. Payoffs are then determined and distributed according to the correspondence between states of the world and actions taken.[7]

Lewis considered games where the sender and receiver had perfect common interest: they both agreed on a unique best action to be taken in each possible state of the world. This need not be the case, however. For discussions of deception, cases of perfect common interest are

[6]The state of the world need not refer to the state of an external world. It can also be interpreted as a private type or quality of the sender.

[7]In some more complicated cases, the message that is sent might also affect the payoffs. One reason for this might be costs associated with production of the message.

not very interesting.[8] In the cases considered below, there will be partial common interest between the sender and receiver. For some states of the world they will agree on a unique best action, but for other states they will disagree.

These games are sometimes presented as an extensive form game tree. An example tree is shown in Figure 1.1. In this example, the sender and receiver have perfect common interest. Play begins with the random choice of nature (N) from the center node. The sender (S) then observes whether she is on the left or right side (whether q_1 or q_2 obtains) and chooses to send message m_1 (up) or m_2 (down). The receiver then observes whether she is at the top (m_1) or bottom (m_2), but not which side she is on (indicated by the dashed lines, which stand for information sets). After choosing an action a_1 or a_2, the sender and receiver then both receive the payoff listed at the end of the actual path taken.

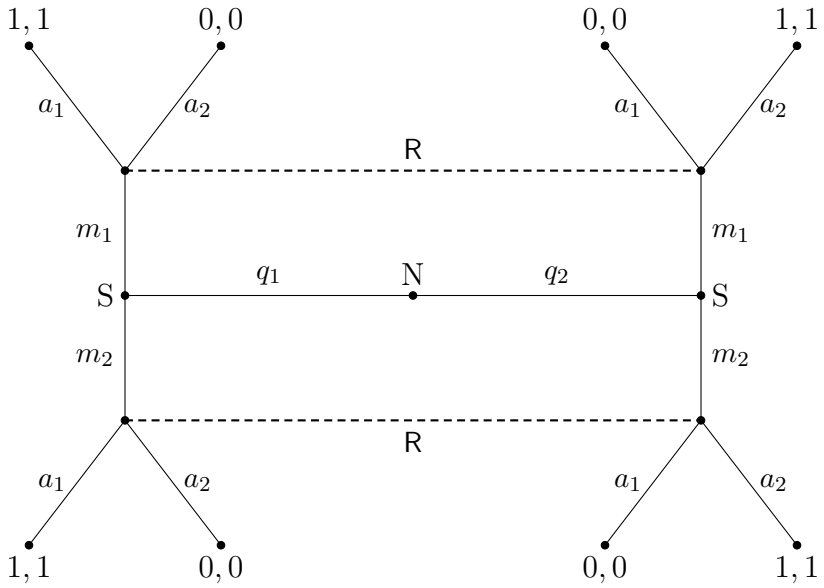


Figure 1.1: Extensive Form Game Tree Example

We can represent these situations precisely and more generally in a mathematical framework. The states of nature can be represented by a set $Q = \{q_1, \ldots, q_n\}$. Each of the states is chosen

[8]This lack of interesting properties is due to the fact that perfect common interest should rule out the "temptation" to be deceptive.

by nature according to some fixed probability distribution. The set of messages available to the sender can be similarly represented by a set $M = \{m_1, \ldots, m_k\}$, and the set of actions available to the receiver can be represented by a set $A = \{a_1, \ldots, a_n\}$. A pure strategy for the sender is a function $S\colon Q \to M$, and a pure strategy for the receiver is a function $R\colon M \to A$. Finally, the payoff functions for a sender and receiver playing pure strategies can be given by payoff functions $\pi_s\colon Q \times M \times A \to \mathbb{R}$ and $\pi_r\colon Q \times M \times A \to \mathbb{R}$.[9]

[9]These payoff functions, as written, include the possibility of signals being costly. If signals are not costly, the payoff functions can be thought of as simply functions of $q \in Q$ and $a \in A$ only. In such cases, they will be written with only two arguments.

Chapter 2

Defining Deception

Having a widely applicable definition can enable effective investigations of some interesting evolutionary questions. For example, Trivers and others have argued that self-deception is evolutionarily advantageous because it aids in other-deception. (see, e.g., Trivers, 2000; von Hippel and Trivers, 2011) Evaluating these claims requires a good definition of what exactly is deceptive in the first place. Standard definitions of deception in human communication involving belief, intent, etc. are inapt for application to these scenarios.

Several attempts have been made to provide a definition relying only on behavioral descriptions. None of them, however, has been entirely adequate for various reasons. These include:

1. not distinguishing between mere (non-beneficial) error and deception;

   Early definitions in particular had this problem. For instance, the first definition offered by Robert Mitchell (1986) does not differentiate mere error from deception. Mitchell himself recognized this as a necessary component, and included additional conditions to include it in a revised definition.

2. inappropriately differentiating providing and withholding information;

   Mitchell's two definitions of deception both of have this problem. He writes, "deception involves providing rather than retaining information." (Mitchell, 1986, p. 17) One need not limit the definition in this way, and in fact, it is probably better not to. There are many situations we would want to call deceptive that involve *not* providing information, such as a chimpanzee not signalling that he or she had found a source of food.

3. insufficient clarity about the meaning of a signal;

   A particular example of this can be seen in the definition from Searcy and Nowicki cited above. The definition they suggest includes the condition of an action being "appropriate if [the signal] Y means X." In the example of Kitui, the "false" alarm call could mean "there is a leopard". It could also mean "I think you should run up a tree". If it means the former, then the situation is deceptive. If it means the latter, then it is not (assuming Kitui "thinks" that the receivers should in fact run up trees). In general, I think that attributing propositional content to signals in non-human animals is probably a mistake altogether, although there are several other positions on this topic as well. (see, e.g., Harms, 2004; Millikan, 2004) I agree with Skyrms that signals having propositional content is not necessary to understand deception and other aspects of signalling, at least most of the time. (Skyrms, 2010, ch. 3)

4. insufficient clarity about benefits and detriments;

   Maynard Smith and Harper's (2003) definition of behavioral deception includes the clause "deception can occur if there is another circumstance in which the same response benefits the signaller at the receiver's expense." It is unclear what counts as an expense in this sense. It is contrasted with receiving a benefit, but does receiving a lesser benefit than one might have otherwise received qualify as an expense, or must the expense be a loss below some fixed baseline?

5. requiring specific structures for the states of the world (mate choice models);

6. discord with intuition in relatively clear cases (Lachmann and Bergstrom); and

7. discord with intuition in cases of uniform population behavior (Skyrms)

The first four problems identified here are typically found in informal definitions of behavioral deception. Formal definitions involving signalling games generally avoid them. The latter three problems, however, are specific to those formal definitions.

Before describing the definition I propose, I will discuss the last three problems listed here in more detail. Although the definition I will propose uses a substantial amount of mathematical formalism, applying that formalism is not the primary focus of this chapter. Some examples will be given, but the main points will be more conceptual in nature.

2.1 Deception as Dishonesty

Problem 5–the necessity of particular structure in the model–is evident in mate choice models and the notions of honesty and dishonesty they employ.[1] In these models, honest signalling occurs when all sender types are uniquely identifiable from the signals sent. Dishonesty occurs when that is not the case.

Definition 2.1 (Dishonesty)**.** Deception occurs when not all states of the world are uniquely identifiable from a signal.

On the surface, this does not appear to even differentiate deception (dishonesty) from mere error. Failure to uniquely identify each state of the world need not be beneficial to the sender. Special structure is built in to the situation to take care of that in many cases. The senders

[1]These models occur in papers by Johnstone and Grafen (1993), Bergstrom and Lachmann (1997, 1998; 1998), Számadó (2000), and Noldeke and Samuelson (2003), among others. The problem I discuss is not an issue for the aims of those papers, but it does hinder more general applicability of this potential definition of deception or dishonesty.

are generally supposed to be of several, linearly ordered quality types, and those types are typically assumed to affect the cost of sending various signals. For each type of sender, the receiver has a unique preferred reaction. Senders, on the other hand, do better (independent of signal costs) if the receiver performs the action preferred for a "higher" type. Thus, any message that does not uniquely identify a sender's type is beneficial to some sender and detrimental to the receiver, separating deceptive signalling from error in that case.

Requiring linearly ordered states of the world is not generally applicable. In the case of Kitui, for example, the signal is not used to indicate a private type of the sender. Instead, it indicates an external state, which is not part of a natural order. Additionally, the signals are not differentially costly depending on the state of the world. Defining deception as dishonesty does not apply well to examples that aren't concerned with mate selection or similar situations.

2.2 Deception as Negative Value of Information

Problem 6–discord with intuitions in relatively clear cases–is illustrated by a definition of deception proposed by Lachmann and Bergstrom (2004):

Definition 2.2 (Lachmann and Bergstrom)**.** A message is deceptive if it has a negative value of information.

The value of information for a message is a statistical measure given by the difference between the receiver's expected payoff when using the message and the expected payoff if the receiver had taken the best possible action without receiving any message. Formally, the value of a message m is given by the expression

$$V_{sig}(m) = \underset{q \in Q_m}{E} \pi_r\left(q, R(m)\right) - \underset{q \in Q_m}{E} \pi_r(q, r_*) \tag{2.1}$$

where $Q_m \subseteq Q$ is the set of states of the world in which the sender sends m, $\pi_r(q, r)$ is the receiver's payoff in world state q for doing action r, and r_* is the receiver's best response if she had not received a signal from the sender.

Lachmann and Bergstrom prove that in a standard signalling game as described in section 1.3, the value of information of any message is guaranteed to be non-negative in equilibrium. So, deception is impossible in equilibrium unless other factors are included in the model. Standard signalling models–without other factors–are often used to describe situations that we would want to call deceptive, however. What about the example of Kitui, the vervet who gave "false" alarm calls to prevent another male from transferring into his troop? The system of vervet alarm calls is an example of a conventional signalling system in nature that is modelled using a standard signalling game. If such a model is appropriate and this definition of deception is correct, then deception should be impossible in equilibrium. Considering ecological cases, it is doubtful that this conclusion is appropriate.

One might argue that nature is hardly ever in equilibrium, so this shouldn't matter. In particular, Kitui might not be playing an equilibrium strategy. So, the use of that call could be considered deceptive. Perhaps this is the case for the particular situation of Kitui. Other situations can be constructed formally that still create issues for this definition. The case of chicks begging for food (modelled with the Sir Philip Sidney game, to be described later) is one such example.

Defining deception as the presence of negative values of information also does not effectively differentiate deception from mere error. The calculations required to identify the value of a signal depend on the payoffs of the receiver alone. Using a signal in a way that induces negative value for the receiver might, in fact, be quite detrimental to the sender.[2] This definition of deception would still identify such a use as deceptive. A sensible definition of

[2]This would only be possible out of equilibrium for two reasons. First, the sender would rather do something better if possible, and second, negative value of information can only occur out of equilibrium.

deception should involve the interests of both the sender and the receiver. Relying only on one, or identifying a situation where the sender and receiver both performed poorly as deceptive, is not ideal.

2.3 Deception as Misinformation with Payoff Conditions

A different formal definition, suggested by Skyrms, employs some tools from information theory to better identify deceptive situations. (Skyrms, 2010, ch. 6) The central tool employed in this definition is the directed divergence measure of Kullback and Leibler.[3] (Kullback and Leibler, 1951) For the context of deceptive behavior in signalling games, there is one specialized case of this measure that Skyrms employs. This is the measure of the amount of information that a message m carries in a particular state of the world q, given the sender behavior S. This is shown in equation (2.2).[4]

$$I(m, q; S) = \log \frac{Pr(q|m; S)}{Pr(q)} \tag{2.2}$$

Skyrms describes this as "the information in [the signal m] in favor of that state [q]."(Skyrms, 2010, p. 36) He also suggests the following intuitive way to think about it. If $I(m, q; S)$ is positive, then receiving the message m makes the state q more likely. Conversely, if $I(m, q; S)$ is negative, then receiving the message makes the state less likely.

Skyrms uses this measure to define the concept of misinformation, a key component of his

[3]This measure can be interpreted as the information in an experimental result, or the mean information in a class of experimental results, for discrimination in favor of one hypothesis against another.

[4]For equation (2.2), the logarithm is conventionally taken to be base-2, measuring information in bits (though that need not be the case). All states q are usually assumed to have $Pr(q) > 0$ so that particular edge case is not an issue. Where the sender behavior S is clear from context, the quantity $I(m, q; S)$ will be written as just $I(m, q)$. This latter form is the notation Skyrms uses.

definition of deception.

Definition 2.3 (Misinformation)**.** Let $\langle Q,\ M,\ A, \pi_s, \pi_r \rangle$ be a signalling game. A message $m \in M$ contains misinformation about a world state $q \in Q$ exactly when either $I(m, q; S) < 0$ or $I(m, q'; S) > 0$ for some state of the world $q' \in Q$ that is distinct from q.

This says a message carries misinformation just in case receiving it either decreases the probability of the actual world state or increases the probability of an incorrect world state.(Skyrms, 2010, pp. 74–5) Deception, according to Skyrms, requires misinformation; but signals that carry misinformation need not be deceptive. The misinformation could be a result of a communication bottleneck, simple error, or some other innocent factor. If, however, the misinformative signal "is systematically sent to the benefit of the sender and the detriment of the receiver, it is *deception*."(*ibid.*, p. 75) Putting these conditions together, Skyrms's definition of deception can be formulated as follows.[5]

Definition 2.4 (Skyrms)**.** Let $\langle Q,\ M,\ A, \pi_s, \pi_r \rangle$ be a signalling game. The use of message $m \in M$ in world state $q \in Q$ by the sender type/strategy S signalling to the receiver type/strategy R is a case of **deception** just in case

i) the use of message m by sender type/strategy S contains misinformation about the state q (definition 2.3);

ii) $\pi_s(q, R(m)) > \pi_s(q, BR_r(q))$ (sender benefit); and

iii) $\pi_r(q, R(m)) < \pi_r(q, BR_r(q))$ (receiver detriment)

Here $BR_r(\cdot)$ is the best-response correspondence for the receiver to the state of the world.

This definition avoids many of the problems of the definitions I have discussed so far. It is possible in equilibrium to observe misinformative messages that benefit the sender at the

[5]This definition is not explicit in Skyrms's work. I have reconstructed it from the chapter referenced above.

expense of the receiver in standard signalling games. So, deception is possible in equilibrium. Furthermore, Skyrms's definition separates deception from mere error, and it does not require any special structure.

This definition is not completely acceptable, however. The reason for this is that uniform behavior in populations should not be considered deceptive. Uniform behavior in this sense means that each agent in the population is behaving in exactly the same way. This corresponds to the population playing a pure strategy in the game.[6] When considering behavior in a signalling game or the behavior of non-human animals, evaluation of actions must be based on what can be seen. There are no intentions or expectations available to include as components of an evaluation.

Skyrms correctly identifies the meaning of a message as a central component of the notion of deception; a message being used deceptively requires a receiver to be misled by the meaning.[7] He is also correct that the meaning of a message is determined by how the message is used. The problem lies in how he understands the use of a message. Skyrms considers the meaning of a message to be determined by the relation of its use to the states of the world. This isn't the right way to think about meaning for the purposes of deception. Instead, the meaning of a message is determined by how the population as a whole uses it.

Suppose an entire population of senders chooses a particular message to represent two different states of the world, say q_1 and q_2 (where there are at least 3 states of the world). Skyrms would call this message a half-truth. In both cases (q_1 and q_2) receiving the message raises the probability of the actual state of the world, but it also raises the probability of an

[6]Ruling out *uniform* behavior as deceptive should not necessarily entail that *universal* deception is impossible. It only rules out one particular kind of potential universal deception–the kind where each agent is acting in exactly the same way. It seems entirely possible that universal deception could still occur in other ways. Determining whether universal deception is actually possible given the definition I propose would require further study that goes beyond the scope of this paper.

[7]This is not unique to Skyrms, by any means. Searcy and Nowicki (2005), Maynard Smith and Harper (2003), and many others all include some criteria involving the meaning of a signal in their definitions of deception.

incorrect state of the world. Contrary to Skyrms, it seems most natural, as Godfrey-Smith (2012, pp. 1295–6) suggests, to say that the meaning of the message is something like a simple disjunction.[8]

Why should that be the case? When interpreting animal signals, one could try to apply Quine's (1960) methodology, though notions of assent and dissent when inquiring about the use of a term would be problematic. Alternatively, one could try to employ a theory like Davidson's (Davidson, 1973). In the case of animals, however, there is no rationality or theory of mind to reliably fall back on to interpret a signal in a way like Davidson suggests. Davidson relies on the interpreter to be able to determine that the speaker intends to convey a truth, in particular, and no method of identifying that kind of intentionality is readily available when considering animals. After removing the requirement of intention to convey a truth, however, a pattern similar to Davidson's still seems to be appropriate.

For concreteness, let's suppose we're concerned with a troop of chimpanzees in whose range is a river and a lake. Observations are made that individuals in a population all use a certain pattern of grunts when they return to the troop having found food near either the lake or the river. In response, other chimpanzees set out to recover more of the food. Some head towards the river, others towards the lake.[9] How should we interpret that pattern of grunts? Some chimpanzees responded inefficiently by going to the wrong location. Was the pattern of grunts deceptive in those cases?

Embracing some core ideas of Davidson (1973, 322ff), I want to argue that the pattern of grunts is not deceptive. Instead, the pattern should be interpreted as something like "there is food near the water" or "there is food near either the lake or river." I don't mean to say that the signal has propositional content necessarily. I am only trying to argue that its

[8]This claim is based on the repetitive context in which the message is sent. Whether this holds or not for the case of one-shot interactions is another matter.

[9]Whether this kind of behavior is actually observed or not in chimpanzees is not important. It is at least possible and is intended to serve as an illustration only.

meaning is not as specific as it could be.

The key to this is the fact that the chimpanzees are all part of the same signalling community who all use the signal in the same way. Although we cannot reasonably say whether the chimpanzees intend to tell the truth (if they have intentions at all), we can still try to generate a best-fit to the observations, trying to interpret them as conveying correct information as often as possible. Davidson's explicit theory asks us to generate "a theory that satisfies the formal constraints on a theory of truth, and that maximizes agreement, in the sense of making [speakers] right, as far as we can tell, as often as possible." (1973, p. 323) Since the troop is uniformly using the signal, the way to maximize the correct transfer of information is to interpret the signal as something like a disjunction.

It is likely impossible to apply Davidson's full theory to all the cases we would want. Even if we could "find a way to interpret the utterances and other behaviour of [a chimpanzee] as revealing a set of beliefs largely consistent and true by our own standards," there is no guarantee that we could do so for all of the situations that are commonly modelled by signalling games. I think the core ideas remain applicable, however. We can't interpret the full meaning of a signal from the chimpanzees; we have no way of determining whether it has propositional, imperative, or has some other kind of content. We can, however, go as far as saying that, whatever the signal happens to mean, all of the creatures in question are using it appropriately if they are all observed to be using it in the same way.

Such disjunctive signals should not be considered deceptive. In any particular instance, receivers have more or less reliable expectations about how a signal is being used. Expectation, in this sense, can be measured by how likely it is that the best response to the population usage of a signal is a best response to how an individual sender is using it in some particular instance. The more likely that is, the more reliable the expectation about the signals use is.

In a situation where senders are each using every signal in the same way, receivers have

perfect expectation about how a signal is being used. This perfect expectation indicates that the receiver "correctly understands" the meaning of any signal being used in a particular interaction as being a disjunctive truth.[10] Therefore, there is no deception in such a case.

This is parallel to the following situation in humans. Suppose that a sender attempts to deceive by telling a half-truth (in Skyrms's sense). If the receiver correctly expects that it is a half-truth and responds appropriately, she is not deceived. If, however, the receiver doesn't expect that it is a half-truth (perhaps because not very many people use the phrase expressing the half-truth in that way), then she might be deceived into taking an inappropriate action.

If this is correct, then no one is using a message in a misleading manner in a uniform population. Receivers have learned to react to how the senders in the population use each message. As that occurs, only usage that differs from what receivers have become accustomed to misleads the receivers in the manner required for there to be deception. Therefore, Skyrms's understanding of the meaning of a message is insufficient. It is not the connection between the use of a message and the world that is important. Rather, it is the connection between the use of a message in one case and how that message is used within the surrounding population that is key.

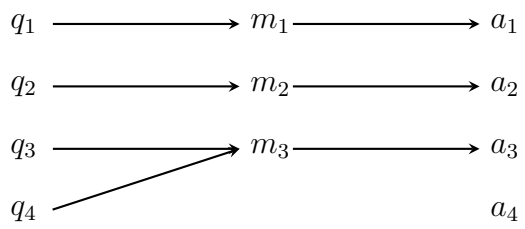
For an example of this, consider a signalling game with $Q = \{q_1, q_2, q_3, q_4\}$, $M = \{m_1, m_2, m_3\}$, $A = \{a_1, a_2, a_3, a_4\}$, with all states of the world equiprobable and the state-act payoffs for the sender and receiver given in the table in Figure 2.1a. One uniform-population equilibrium is given by the behavior depicted in Figure 2.1b. This equilibrium corresponds to a situation in which there is a uniform population of senders, each performing the described behaviors.

Skyrms's definition indicates that the use of message m_3 in this equilibrium is deceptive. We

[10] Again, the receiver need not consciously identify the message as being disjunctive, as being a truth, or anything of that sort. The "correct understanding" is indicated by the appropriateness of the response taken.

	a_1	a_2	a_3	a_4
q_1	$1, 1$	$0, 0$	$0, 0$	$0, 0$
q_2	$0, 0$	$1, 1$	$0, 0$	$0, 0$
q_3	$0, 0$	$0, 0$	$1, 1$	$0, 0$
q_4	$0, 0$	$0, 0$	$1, 1$	$\frac{1}{2}, \frac{3}{2}$

(a) State-Act Payoffs



(b) Equilibrium Behavior

Figure 2.1: Signalling Game with an Information Bottleneck

can calculate that message m_3 is misinformative when the state of the world is q_3 or q_4:

$$I(m_3, q_3) = \log \frac{Pr(q_3|m_3; S)}{Pr(q_3)} = \log \frac{Pr(q_4|m_3; S)}{Pr(q_4)} = I(m_3, q_4) = \log \frac{0.5}{0.25} = 1 \tag{2.3}$$

But just the presence of a misinformative message does not guarantee deception. Skyrms also requires that there is a systematic sender benefit and receiver detriment. In world state q_4, these requirements are met. In world state q_4, the sender earns a payoff of 1, whereas if the receiver knew that the state of the world was indeed q_4, the sender would only earn $\frac{1}{2}$. Additionally, in world state q_4, the receiver earns a payoff of 1, whereas if the receiver knew that the state of the world was indeed q_4, she could earn $\frac{3}{2}$ instead. Therefore, under Skyrms's definition of deception, the message m_3 is used deceptively in world state q_4.

This is not a correct identification. The game considered has an information bottleneck; *there aren't sufficiently many signals available to uniquely identify all states of the world.* The senders, at least in this case, are all uniquely identifying as many states of the world as they possibly could. More importantly, all senders are sending the same messages in the same states. Although there is misinformation in Skyrms's sense, the meaning of the message m_3 is unanimous. This seems to be a natural case where we should say that the half-truth of the message corresponds to something like a simple disjunction over the states of the world. No sender is deviating from the established pattern of usage of the message, and thus the use of the message should not be considered deceptive.

2.4 A Revised Definition

Another way to see the central problem both with defining deception as a message having negative value of information and with defining deception based on misinformation as Skyrms suggests is that the factors influencing deceptive behavior are not completely specified. In the case of deception as negative value of information, the unit of analysis was a particular message, given some sender behavior: a (message, sender, receiver) combination. In Skyrms's case, the thing that is deceptive is a particular message, given some state of the world and some sender and receiver behaviors: a (message, state, sender, receiver) combination. I want to suggest that neither of these is the correct way to think about deception. Instead, a (message, state, sender, receiver, population) combination is the kind of thing that is deceptive. The difference here is the added context of the surrounding population. The population as a whole determines what a signal "means" through using it in particular ways, and only by considering that can we correctly understand deceptive signals. The question is how to integrate the information about the population into a definition of deception.[11]

Integrating this thought into a formal framework can be accomplished by a relatively straightforward alteration of Skyrms's definition of deception. This alteration replaces misinformation with an alternative concept: misuse. Instead of comparing the posterior probabilities of states given the use of a signal and the prior probabilities of states, the correct thing to do is to compare the posterior probabilities of states given the use of a signal by a particular type with the posterior probabilities given by the use from an average member of the population. By an average member of the population, I mean the behavioral mixed strategy σ_P determined by the composition of the pure strategies in the state of the population P.[12] Making

[11] The method of integration I will propose treats sender behavior and the average behavior of the population separately. This is potentially related to Godfrey-Smith's suggestion that "There is a difference between the *maintaining* and the *non-maintaining* uses of the signal. Some uses contribute to stabilization of the sender-receiver configuration and some, if more common, would undermine it. Those ones are deceptive." (Godfrey-Smith, 2012, p. 1295) Godfrey-Smith and I seem to agree that thinking about deception requires considering both particular usage of a signal and the usage of a signal in the population as a whole.

[12] I do not mean the modal pure strategy or any other possible interpretation. It is not necessary that

the formal adjustments results in the following measure, analogous to equation (2.2).

$$J(m, q; S, \sigma_P) = \begin{cases} \log \frac{Pr(q|m;S)}{Pr(q|m;\sigma_P)} & \text{if } Pr(q|m;\sigma_P) \neq 0 \\ 0 & \text{otherwise} \end{cases} \tag{2.4}$$

Skyrms described the formula in equation (2.2) as "the information in [the signal m] in favor of that state [q]."(Skyrms, 2010, p. 36) This new formula can be thought of as the information in m in favor of q beyond (or short of) what the population average would indicate. Given this new measure, we can define misuse of a signal.

Definition 2.5 (Misuse)**.** Let $\langle Q,\ M,\ A, \pi_s, \pi_r \rangle$ be a signalling game. Sender type S **misuses** a message $m \in M$ in the world state $q \in Q$ relative to the population P exactly when

(i) $Pr(m|q; S) > 0$;

(ii) $J(m, q; S, \sigma_P) > 0$; and

(iii) $J(m, q'; S, \sigma_P) < 0$ for some world state $q' \in Q$ with $q' \neq q$

The first requirement in the definition is that the message is actually used in the world state q by S with some positive probability. The second and third conditions do the real work. The second condition requires that the real state of the world q be more likely after receiving the message m from S than from an average member of the population. If the receiver knew the sender's type/strategy, she could more reliably identify the state of the world. The third condition requires that receiving the message m from an average member of the population makes a false state q' more likely than receiving the message from S would. Since it is assumed that the receiver cannot reliably identify the type/strategy of the sender, she has

anyone in the population actually be playing σ_P as their strategy. In fact, in many models, that would be impossible when the population is polymorphic.

to respond as if to an average member of the population. These conditions mean that the sender type S is 'hiding' in the population; the information in the sender's message is being obscured through the receiver's response to the population state.

Given this notion of misuse, I propose the following revision of Skyrms's definition of deception.

Definition 2.6. Let $\langle Q,\ M,\ A, \pi_s, \pi_r \rangle$ be a signalling game. The use of message $m \in M$ in world state $q \in Q$ by the sender type/strategy S signalling to the receiver type/strategy R given surrounding population P is a case of **deception** just in case

i) the message m is misused by sender type S in state q relative to population P (definition 2.5);

ii) $\pi_s(q, R(m)) > \pi_s(q, BR_r(q))$ (sender benefit); and

iii) $\pi_r(q, R(m)) < \pi_r(q, BR_r(q))$ (receiver detriment)

Here $BR_r(\cdot)$ is the best-response correspondence for the receiver to the state of the world.

The measure of misuse of a signal relies on the existence of a population within which an individual or a strategy type is embedded. This definition does not require that the population be in equilibrium (or out of equilibrium), nor does it even require that there are any particular dynamics governing its evolution. It only requires that one is able to specify the other individuals who are involved in the population and the type/strategy of each. If an infinite population is being discussed, the relative frequencies of the various types/strategies in the population need to be specified. Whether this definition *can* be applied in various situations is not the same as whether it is a good definition. So, I will consider an example of a typical model where deception or dishonesty is expected and see how this revised definition fares.

2.4.1 Begging behavior and the Sir Philip Sidney game

Huttegger and Zollman (2010) have analyzed a scenario where deception should be expected. For some equilibria of this scenario, Skyrms's definition of deception and the one I have proposed agree that deception is present. In an equilibrium of a slightly expanded game, however, the definitions disagree.

Huttegger and Zollman analyze the Sir Philip Sidney game (Maynard Smith, 1991) as a model of chicks begging for food from a parent. Chicks can be in one of two states–Needy or Healthy–and they can choose to either beg for food or not. Parents, having observed the begging or lack thereof, can choose to donate a portion of their own resources to the chick or not. Without a donation, Needy chicks are less likely to survive than Healthy ones. If a chick receives a donation, then it is more likely to survive than if it didn't, and with a donation of resources, both kinds of chicks are equally likely to survive. If a parent donates some of its resources, then it is less likely to survive than if it did not donate.

Thinking only of the ecological analogue of the formal model, we might expect begging to indicate that a chick is Needy.[13] We might also expect some chicks to be deceptive by always begging for food, whether they are Needy or not.

The standard equilibria identified for the Sir Philip Sidney game all have uniform populations–all senders (chicks) are behaving identically, and all receivers (parents) are behaving identically.[14] Interestingly, though, in addition to the standard uniform population equilibria, Huttegger and Zollman find that under certain parameter conditions, there can be a hybrid equilibrium. In this equilibrium, some chicks beg only when they are Needy while others always beg, and some parents donate only if there is begging and others never donate.[15] In

[13]Begging might alternatively indicate health instead of neediness, but that isn't what we see in nature generally.

[14]The precise details of how they behave might vary among different possible equilibria, but within each equilibrium, the populations are uniform in their strategies.

[15]Following Huttegger and Zollman in representing the sender strategy that begs only when Needy as S_2

this case, Skyrms's definition and mine agree that there is indeed deception.

In this hybrid equilibrium the average sender population strategy σ_P is the strategy that begs with probability 1 if its state is Needy and begs with probability $(1-\lambda)$ if it is Healthy. This behavior is shown in Figure 2.2.[16]

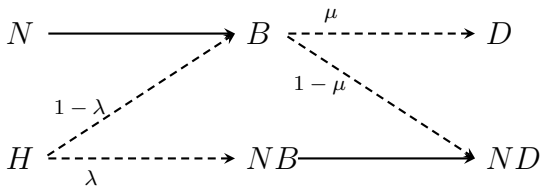


Figure 2.2: Hybrid Equilibrium Behavior Map

Chicks that beg when they are Healthy are identified as deceptive by both my definition and Skyrms's. In the case of my definition, the signal "beg" (B) is mis-used relative to the hybrid population in each state by senders who always beg.[17] Furthermore, this misuse of the signal "beg" is deceptive when the chick is Healthy, as it also exhibits sender benefit and receiver detriment. If the sender is Healthy, the best response for a receiver if she knew that fact is to not donate resources.[18] In an actual interaction, though, these senders induce

and the strategy that always begs as S_4, the hybrid sender population is a convex linear combination of those strategies. The precise combination is $\lambda S_2 + (1-\lambda)S_4$, where $\lambda = \frac{k(ma+(1-m)b)-d}{(1-m)(kb-d)}$. (equation (2.6) in Huttegger and Zollman, 2010) Here, k is the relatedness coefficient, b is the benefit from a donation to a Healthy chick, and d is the percentage of the resource the parent donates. Additionally, a is the benefit from a donation to a Needy chick and m is the probability that an arbitrary chick is Needy.

[16]The value $\mu \in [0,1]$ is defined similarly to λ in equation (2.6) of (*ibid.*). The precise definition is unimportant here, though.

[17] In particular, the relevant values are as follows, where m is the prior probability of a chick being Needy:

$$Pr(B|H;S_4) = 1 > 0 \tag{2.5}$$

$$J(B,H;S_4,\sigma_P) = \log\left(\frac{m}{1-\lambda} + (1-m)\right) > 0 \tag{2.6}$$

$$J(B,N;S_4,\sigma_P) = \log\left(1-\lambda+\lambda m\right) < 0 \tag{2.7}$$

Noting that $P(B|N;S_4) = 1$, we can also see from these calculations that begging by S_4 when the type is Needy is *not* misusing the signal.

It is worth pointing out that the sender type S_2–the type that begs just in case it is Needy–also misuses the signal "beg" relative to the population when the chick is Needy, instead of Healthy. This misuse is not deceptive, however, as there is no benefit to the sender.

[18]The receiver would donate if $b > \frac{d}{k}$ and not donate if $b < \frac{d}{k}$. The requirements on the existence of the hybrid equilibrium given by Huttegger and Zollman include this second inequality.

many receivers into donating, so there is receiver detriment. There is also benefit to the sender relative to the receiver knowing the true state.[19] Therefore, these senders' begging when Healthy in the hybrid equilibrium is deceptive to receivers who donate.

If we expand the Sir Philip Sidney game slightly, we can also see a situation where Skyrms's and my proposed definition disagree. Instead of having only two types of chicks, suppose there are three: Needy, Moderately Healthy, and Healthy. We can also adjust the available signals to include a message that could theoretically allow perfect information transfer: Begging, Weak Begging, and No Begging. The payoffs for the game are modified to account for this by asserting that the chance of a Moderately Healthy chick surviving without receiving a donation is between that of the Needy chick and the Healthy chick.[20]

In this expanded game, there is some behavior that Skyrms's definition identifies as deceptive but the one I have proposed does not. One example of this is given in Figure 2.3.[21] Skyrms's definition identifies the Begging behavior of the Moderately Healthy chick as deceptive, given certain parameter values for the game.[22] The reason Skyrms and I disagree on this example is that there is no misuse of signals. The meaning of the signals, as established by the population of chicks, is uniform, and it is therefore not exploited by any individuals or types.

[19] In an actual interaction, the sender has a payoff of $(1-c)+k(1-d) = 1+k-c-kd$ instead of $(1-b-c)+k = 1+k-c-b$ if the receiver knew the true state. The first of these quantities is greater than the second so long as $kd < b$ or $b-kd > 0$. Yet another of the requirements on the existence of the hybrid equilibrium require that $b-kd > c \geq 0$, so this condition holds. Thus, there is a benefit to the sender.

[20] In the original game, the probability of survival without a donation for a Needy chick is $1-a$ and the probability of survival without donation for a Healthy chick is $1-b$, where $1-a < 1-b$. With the addition of a new type, we can add a new parameter e so that $1-e$ is the probability of a Moderately Healthy chick surviving without a donation, where $1-a < 1-e < 1-b$.

[21] This is not necessarily equilibrium behavior.

[22] The particular parameter values that are required are that $d > ke$ and $e > kd$, where d is the proportion of the resource donated by the parent, k is the relatedness coefficient between the parent and chick, and e is the benefit of a donation when it is received by a Moderately Healthy chick. These are required in order for there to be sender benefit and receiver detriment in the case where the chick is Moderately Healthy.

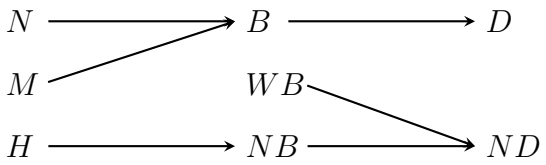


Figure 2.3: Example Behavior in the Expanded Sir Philip Sidney Game

2.5 Stomatopod Bluffing

As a second example showing deception without regard for whether a system is in equilibrium, I will consider a sketch of a model of stomatopod bluffing behavior.[23] Stomatopods make threat displays to conspecifics to deter attempts to take a resource without actually having to fight. Periodically, stomatopods also molt, losing their ability to fight for a short time as the exoskeleton regrows. During these molting periods, however, threat displays are still observed. If the molting individual is challenged, the stomatopod flees instead of actually fighting. These are the threat displays that are typically identified as deceptive bluffing.

A rough model of this scenario can be constructed as follows. There are three possible states of the world: molting and not willing to fight ($M.NF$), not molting and willing to fight ($NM.F$), and not molting and not willing to fight ($NM.NF$).[24] In each state of the world, a sender has two options: make a threat display (D) or don't (ND). A receiver can then observe whether or not there was a threat display and choose to challenge for the resource (C) or not (NC). If the receiver does not challenge for the resource, the sender retains it. If the sender is willing to fight and the receiver challenges, a (costly) fight ensues. If the sender is not willing to fight but the receiver challenges, the sender flees and the receiver takes the resource. Finally, in all other cases, the sender retains the resource.

[23]For further information on the actual bluffing behavior, see, e.g., Steger and Caldwell (1983) or Searcy and Nowicki (2005, pp. 160–169). The following model is not intended to be a good approximation of realistic behaviors. It is intended as a rough approximation in order to consider whether similar behavior is deceptive according to the definition I have proposed.

[24]One might also include a fourth state – molting and willing to fight ($M.F$) – but since I am not concerned with dynamic behavior here, I will omit it for the sake of clarity.

The specifics of the payoffs in this model are not important. I am not concerned with equilibrium behavior in this case. What is important in order to identify whether there is deception or not is the prior probabilities of the states and the composition of the population. As an example, suppose the prior probabilities of the various states of the world are[25]

$$Pr(M.NF) = 0.04$$
$$Pr(NM.F) = 0.48$$
$$Pr(NM.NF) = 0.48$$

To establish the presence of deceptive behavior, we don't need to assume anything about the representation of receiver strategies in the population as long as the following strategy has non-zero representation:

R_2: challenge if there is no threat display, don't challenge if there is

Finally, consider a sender population comprised of the two strategies

S_2: provide a threat display if willing to fight or molting, don't otherwise

S_6: provide a threat display if willing to fight, don't if not (regardless of molting status)

in the proportions

$$Pr(S_2) = 0.5$$
$$Pr(S_6) = 0.5$$

The average population behavior is given in Figure 2.4.

[25]The probabilities were produced keeping in mind the data on length of time between molts from Reaka (1979) and the length of time during a molt from Steger and Caldwell (1983). Although the two data sets are from different species, the numbers seem reasonable as a very rough approximation. The following results don't necessarily hold for all possible priors probabilities that might be specified.

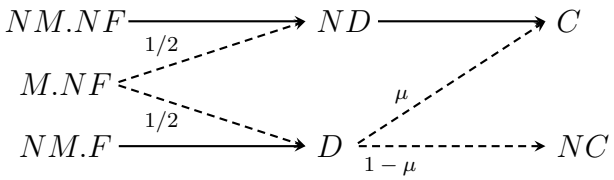


Figure 2.4: Stomatopod Population Behavior Map

It is straightforward to calculate that the use of a threat display by sender S_2 when molting is a misuse of the signal relative to this population.

$$Pr(D|M.NF) = 1 > 0 \tag{2.8}$$

$$J(D, M.NF; S_2, \sigma_P) = \log\left(\frac{25}{13}\right) > 0 \tag{2.9}$$

$$J(D, NM.F; S_2, \sigma_P) = \log\left(\frac{25}{26}\right) < 0 \tag{2.10}$$

This misuse is deceptive with respect to receiver type R_2. If the receiver knew the real state – that the sender was molting – the best response would be to challenge for the resource. In an actual interaction, however, the receiver's response would be to not challenge since there was a display. Given that the sender is molting, the receiver would prefer challenging to not challenging, as the sender is unable to effectively fight. On the other hand, the sender would prefer to keep the resource instead of having to run away. Thus, there are both sender benefit and receiver detriment in this example, satisfying the definition of deception I propose.[26]

[26] There are other misuses of signals in various states by various types in the sender population, but no others are deceptive.For example, sender type S_2 also misuses the signal of not making a threat display (ND) in the state where it is not molting and not willing to fight ($NM.NF$), but this is not deceptive. The receiver's behavior on getting no threat display does not differ from the optimal response to a sender who would run away if challenged. So there is no sender benefit or receiver detriment.

2.6 One-on-one Interactions

The definition of deception I have proposed nominally requires a population of interacting agents. However, it is also applicable to one-on-one interactions. In one-on-one interactions, sender and receiver populations could naturally be identified as each being comprised of exactly one individual. If the sender and receiver populations are both comprised of a single individual, the definition I propose would seem to say that no behaviors could be deceptive. For, both populations would exhibit uniform behavior. In repeated interactions, however, an alternate identification of the populations could be employed.[27]

In repeated interactions, there is a history of play between the individuals. This history can be used to construct historical sender and receiver populations, which would enable the application of my definition. For example, suppose we are interested in two baboons who had been isolated on an island since birth. Let us call them Sandy and Roger. Suppose that in some sort of signalling interaction Sandy has always been the signaller and Roger has always been the receiver. Instead of identifying the sender population as having size one–being comprised of just Sandy–we could consider the historical behavior of Sandy.[28] This would be formally equivalent to the sender population being comprised of many individuals. The strategy distribution would correspond to how Sandy has used messages in the past.[29] Then the definition I have proposed could be readily applied, taking Sandy's behavior at a particular time as the sender type in question for the purposes of inquiring whether the use

[27] In a one-shot interaction, no messages can be deceptive. That is not particularly problematic. In a one-shot interaction, the messages have no meaning, either inherently or as defined by use, before the interaction occurs. Most cases where we would want to say that one-on-one interactions were deceptive are not one-shot interactions.

[28] In this procedure, it might also be pointed out that the meanings of messages change over time, and considering the entire history of use for determining σ_P is inappropriate. Such a suggestion is well taken, and there is a clear response. Instead of considering the entire history of usage, one might institute a finite time horizon, restricting the history to some relevant recent time interval (representing the window of available memory in the individuals, perhaps).

[29] In fact, it isn't even necessary for the purposes of the definition I have proposed to know the distribution of sender strategies in the population. Instead, it is enough to be able to calculate σ_P, which is even more easily determined.

of a message is deceptive.

2.7 Lucky Errors

One of the problems I pointed out at the start was that some definitions of deception do not differentiate deception from error. The definition I have proposed does not have this issue in general. It can, however, identify "lucky" errors (errors that are beneficial to the sender and detrimental to the receiver in the right way) as deceptive. I do not consider this a problem. Whether it does or doesn't pick out "lucky" errors as deceptive is an artifact of the chosen model, which turns the question to which model is most appropriate for the problem at hand. For any particular model, lucky errors are straightforwardly expunged or incorporated, depending on the explanatory aims.

Inclusion of error in a signalling model can be accomplished in (at least) four ways. There can be explicit error terms in perception of the state of the world, in transmission of the signal, in receiving the signal, and in performance of an action. In some of these cases, the errors are irrelevant to the proper application of the definition I have proposed. In others, the errors can be either incorporated or ignored depending on the situation at hand.

Errors in action performance are not important for identifying deception. The definition should be applied using the action that the receiver attempted to perform. Similarly, explicit errors of message perception should also be ignored. Instead, the focus should be on what the receiver would have done if she had received the message correctly.

Errors in perception of the state of the world and in selection of a message to send are different. In these cases there are two options. One could incorporate the error rates into the sender strategy, effectively transforming any pure strategy into some mixed one, and apply the definition to that. Alternatively, one could ignore the error rates and consider

what would have happened in the absence of error. The choice of which of these alternatives to employ depends on the explanatory aims at hand. In either case, the error is modelled explicitly, and can be separated from the sender behavior if desired.

2.8 Discussion and Conclusions

I have so far proposed a new definition for behavioral deception and showed how to apply it in certain cases. It would be flawed, however, if it fell victim to the same problems I have outlined for previous definitions. Fortunately, it does not.

First, the definition I propose does differentiate deception from mere error, as many others do.[30] This is accomplished by requiring sender benefit and receiver detriment. A sender can clearly use signals erroneously, perhaps due to a misidentification of the actual state of the world or any number of other factors. It is only when that misuse is of actual benefit to the sender (and detrimental to the receiver) that the definition I propose would identify it as deceptive.

Second, the signalling game framework easily incorporates deception both by providing and withholding information. Withholding information (by means of not providing a signal) can simply be modelled as another available message in the game. The extent to which the definition I propose avoids or commits this error is then a matter of modelling the situation under investigation appropriately. The definition itself does not rule out either option.

Third, the signalling game framework allows my definition to avoid having to explicitly determine the meaning of a signal. In a standard signalling game, messages begin with no pre-established meaning because their primary use is often to explain how meaning can arise from the strategic interactions required to play the game. This is a significant advantage

[30]This claim is subject to the discussion on lucky errors in section 2.7.

in understanding deception. For if the messages did have established meanings, intentions related to meaning would have to be incorporated. Adding intention would then cause problems in trying to understand deceptive behavior in non-human animals. The use of a message by the population is what matters. The meaning need not be indicative or imperative. It need not even correspond directly to any common characteristic of human language. The only feature the proposed definition considers is whether a message is being used in the same way by the sender as it is being used by the relevant population.

Fourth, the extent to which the definition I propose correctly takes into account benefits and detriments is again a matter of modelling. Given any particular model, there is no ambiguity in whether the use of a message is beneficial or detrimental.

Fifth, the definition I have proposed does not require that any particular structures be placed on the states of the world. The world states might be linearly ordered as would be expected in mate choice models, but they might not be. For example, the relevant states might be the presence of various types of predators, all of which are equally dangerous. No natural ordering is apparent in such a situation. The definition I have proposed is equally applicable in either case.

Finally, the definition I have proposed seems to be in accord with intuition in relatively clear cases. The example of the Sir Philip Sidney game shows how a message could be deceptive in equilibrium in a standard signalling game, contrary to the conclusion arrived at by applying Lachmann and Bergstrom's definition.

It might also be useful to summarize what types of situations can and cannot play host to deceptive behavior, according to the definition I have proposed.

1. polymorphic populations in (mixed) equilibria;

2. polymorphic populations out of equilibrium;

3. idiosyncratic individuals in a nearly uniform population; and

4. repeated individual interactions.

This list is not necessarily exhaustive. One notable instance that cannot play host to deception is the case of a uniform population (either in or out of equilibrium).

In summary, I have proposed a definition of deception in formal models and have argued that it is superior to several prominent alternatives as a definition of behavioral deception. The question of how far the range of applicability of this definition reaches is an interesting question. One might inquire whether it is applicable in all animal species (perhaps some animals have human-like mental states, making an intentional definition more apt) or whether it could be applicable even to humans despite the general tendencies to use intentional definitions of deception in the human case. This remains an open question.

Chapter 3

On the Possibility of Universal Deception

3.1 Introduction

A famous claim often attributed to Kant is that universal deception is impossible. Particular sources of this are rather widespread and varied. One of the more prominent locations for this claim is Kant's discussion in the *Groundwork of the Metaphysics of Morals* of whether it is possible to will that getting out of difficulties by making lying promises should be a universal law. Kant writes:

> Let the question be, for example: may I, when hard pressed, make a promise with the intention not to keep it?... I must reflect whether the matter might be handled *more prudently* by proceeding on a general maxim and making it a habit to promise nothing except with the intention of keeping it. But it is soon clear to me that such a maxim will still be based only on results feared.... [I]f I am unfaithful to my maxim of prudence, this can sometimes be very advantageous

> to me, although it is certainly safer to abide by it. However, to inform myself in the shortest and yet infallible way about this answer to this problem, whether a lying promise is in conformity with duty, I ask myself: would I indeed be content that my maxim (to get myself out of difficulties by a false promise) should hold as a universal law (for myself as well as for others)?... I soon become aware that I could indeed will the lie, but by no means a universal law to lie; for in accordance with such a law there would properly be no promises at all, since it would be futile to avow my will with regard to my future actions to others who would not believe this avowal or, if they rashly did so, would pay me back in like coin; and thus my maxim, as soon as it were made a universal law, would have to destroy itself. (Ak 4:402–3)

Kant mentions a similar situation in the *Critique of Practical Reason*. (Ak 5:21)

Also in the *Critique of Practical Reason*, Kant mentions other situations with the same theme. If there were a universal law that one could deny bank deposits when no proof exists, "such a principle, as a law, would annihilate itself since it would bring it about that there would be no deposits at all." (Ak 5:27) Kant predicts similar effects if everyone were to lie when giving testimony. He writes, "It is obvious that... everyone would be necessitated to truthfulness. For it cannot hold with the universality of a law of nature that statements should be allowed as proof and yet be intentionally untrue." (Ak 5:44). Finally, in "On a supposed right to lie from philanthropy," Kant says that universal deception is impossible for rational agents even more clearly: "Truthfulness in statements that one cannot avoid is a human being's duty to everyone, however great the disadvantage to him or to another that may result from it...." (Ak 8:426) Since honesty is a duty for all humans, then, and duties result from considerations of rationality, universal deception should be impossible.

In all of these examples, there is a common theme. If everyone were to lie whenever it is in their interest, the system in question would destroy itself. Promises would not be believed, deposits would not be made, and testimony would be worthless.

Whether these are accurate depictions of Kant's views or not, many have interpreted them as saying that universal deception is impossible. Skyrms has suggested that Kant is wrong on this count. He provides an example of a signalling game in equilibrium that, according to his definition, exhibits universal deception. He writes, "Every signal sent in this equilibrium is deceptive. Universal deception in this strong sense is not only *logically* consistent in the sense of involving no contradiction, but also *evolutionarily* consistent in the sense of being an equilibrium." (Skyrms, 2010, p. 82)

Although I disagree with Skyrms's definition of deception in signalling games, I do agree with this final conclusion. It is possible to have universal deception, in a very strong sense, in equilibrium in a signalling game. Before turning to examples, however, it is important to make clear exactly what is meant by universal deception.

3.2 Kinds of Universal Deception

In the context of a signalling game, there are several senses of universal deception one might consider. Most of these will be unimportant for evaluating Kant's claim. However, they are logically possible and should be distinguished from the strong sense that is really under consideration. Each is also accompanied by the quantifier structure it exhibits.

1. For each message that is used, there is some state of the world in which it is used deceptively by some sender. ($\forall m \exists q \exists S$)

2. For each message that is used, in every state of the world in which it is used, some

sender uses it deceptively. ($\forall m \forall q \exists S$)

3. For each sender, there is a state of the world in which they use a message deceptively. ($\forall S \exists q \exists m$)

4. For each sender and each state of the world, some message the sender uses is used deceptively. ($\forall S \forall q \exists m$)

5. For each sender and each state of the world, any message the sender uses is used deceptively. ($\forall S \forall q \forall m$)

The first of these can be thought of as universally deceptive in the sense that each message is used deceptively sometimes. This is perhaps the weakest sense on universal deception that one might contrive. The possibility of the analogous notion in human communication would be nearly trivial to confirm. If something could be said as a lie, it likely has been in some state of the world by someone.

The second kind is universally deceptive in the sense that each message is used deceptively by some sender in all the states that it is used in. This is a slightly stronger, but still quite weak notion of universal deception.

The third sense is universally deceptive in the sense that each sender is sometimes deceptive. Unlike the previous two notions, this now focuses on the universality of the deceivers instead of the deceptive messages. In human terms, it is analogous to everyone lying sometimes.

The fourth and fifth senses, in the context of an evolutionary signalling game where senders use only pure strategies are in fact equivalent. For, in each state of the world, a sender using a pure strategy only ever sends one message. These senses are universally deceptive in the strongest sense: each sender is always deceptive.

This last notion, I think, is the closest notion of universal deception to what Kant intended in many of his discussions. He claimed that if everyone would lie when it was in his or her interest, the system would collapse. Although each discussion had a limited scope – testimony, bank deposits, making promises, etc. – the underlying message was always about the universality of lying and deception.

The kinds of universal deception I've proposed don't, in fact, refer to the interests of the deceiver at all. So, how could they be used to understand Kant's claims? The reason for this is that the definitions of deception that Skyrms and I propose build in the condition that a deception should be in the interests of the deceiver. This indicates that perhaps our understanding of deception is not entirely analogous to Kant's notion of lying. But, any disanalogy should be harmless for proceeding to study the possibility of universal deception. Kant is interested in universal lying when it is in the interest of the liar, and this is very much related to the notions of deception that Skyrms and I employ.

3.3 Examples of Universal Deception

3.3.1 Skyrms's Universal Deception

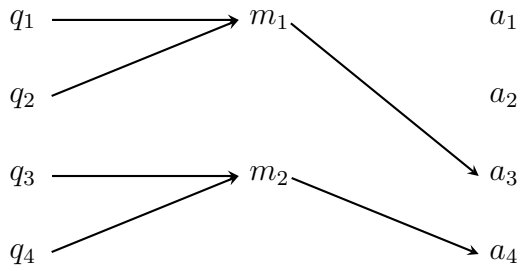
Skyrms's example of universal deception (according to his definition) is of the fifth type. In the equilibrium presented, all senders are being deceptive in every state of the world with any message they send. The relevant state-act payoffs and the equilibrium behavior Skyrms describes are detailed in Figure 3.1.[1]

This behavior is universally deceptive for Skyrms because each message contains what he calls misinformation (see definition 2.3 in chapter 2) that is beneficial for the sender and

[1]The notation differs slightly from Skyrms's own. For the original, see (Skyrms, 2010, p. 81).

	a_1	a_2	a_3	a_4
q_1	$2, 10$	$0, 0$	$10, 8$	$0, 0$
q_2	$0, 0$	$2, 10$	$10, 8$	$0, 0$
q_3	$2, 10$	$0, 0$	$0, 0$	$10, 8$
q_4	$0, 0$	$2, 10$	$0, 0$	$10, 8$

(a) State-Act Payoffs



(b) Equilibrium Behavior

Figure 3.1: Skyrms's Universal Deception Equilibrium

detrimental for the receiver, compared to what would happen if the receiver knew the actual state of the world. That the messages contain misinformation is due to the fact that receiving them raises the probability of a non-actual state of the world along with raising the probability of the actual state of the world.

This is beneficial for the sender because the best response given that information is for the receiver to take act a_3 or a_4, depending on the message. With those responses, the sender earns a payoff of 10. If the receiver knew the actual state of the world, however, she would choose a_1 or a_2 as an action, earning the sender a payoff of only 2.

Acting based on the message is also detrimental to the receiver. If she knew the actual state of the world, she could respond with a_1 or a_2 and earn a payoff of 10. Since she only can determine with certainty that a disjunction of two states is actual after receiving a message, the best response earns a payoff of only 8.

According to the definition I have proposed and defended, however, this behavior is not deceptive at all. The senders are all acting in the same way. So, the meaning of each message is well-established in the population. Message m_1 can be thought of as meaning "State q_1 or q_2 is actual," and message m_2 can be thought of as meaning "State q_3 or q_4 is actual."[2]

[2]These paraphrases should not be construed as accepting a propositional theory of meaning for messages in signalling games. They are simply convenient possibilities to consider.

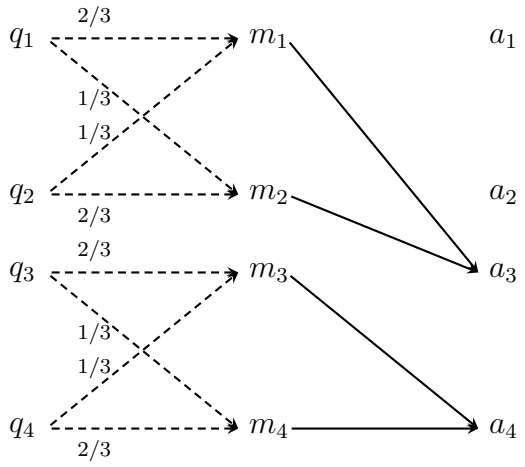
The goal of the following two examples is to explore whether universal deception, in the strong sense that Skyrms was attempting to display, is possible under my definition of deception. First, I will present an example that is structurally similar to Skyrms's own. This will show a strong kind of universal deception, but not the strongest kind that we are after. Simplifying the example, however, will demonstrate that the strongest kind of universal deception is in fact possible in equilibrium.

3.3.2 Nearly Universal Deception

This first example revolves around the state-act payoff table in Figure 3.2a. There are four possible states of the world, four possible messages, and four possible actions. The receiver would like to be able to uniquely identify the actual state of the world and respond with a_1 or a_2. For the senders, however, that is not an optimal response and they can coordinate to not allow that unique identification. One possible set of equilibrium behavior is diagrammed in Figure 3.2b.

	a_1	a_2	a_3	a_4
q_1	$5, 12$	$0, 0$	$10, 10$	$0, 0$
q_2	$0, 0$	$5, 12$	$10, 10$	$0, 0$
q_3	$5, 12$	$0, 0$	$0, 0$	$10, 10$
q_4	$0, 0$	$5, 12$	$0, 0$	$10, 10$

(a) State-Act Payoffs

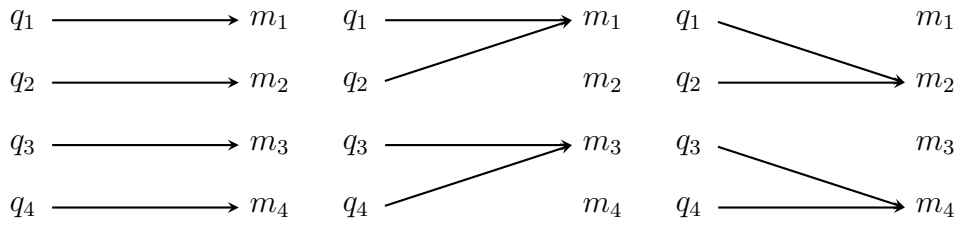


(b) Equilibrium Behavior

Figure 3.2: Nearly Universal Deception Example

There are many ways to reach this aggregate behavior, but the one I want to focus on is a

population comprised of equal proportions of three types: S_1, S_2, and S_3. These types are illustrated in Figure 3.3. Given that sender population, the best response for the receiver is to do action a_3 when receiving either m_1 or m_2, and to do action a_4 when receiving message m_3 or m_4. Similarly, given the receivers' behavior, no sender would benefit by deviating from her current strategy.



(a) Type S_1 (b) Type S_2 (c) Type S_3

Figure 3.3: Nearly Universal Deception Sender Types

In this example, each of the three sender types behaves deceptively in some situations. This shows the possibility of nearly universal deception: each sender is deceptive sometimes (type 3). Sender types S_2 and S_3 behave deceptively in some states of the world, but not all of them. In particular, when senders of type S_2 use message m_1 in state q_1 or m_3 in q_3, or when senders of type S_3 use message m_2 in q_2 or m_4 in q_4 the behavior is *not* deceptive. Although there is a benefit for the sender and detriment for the receiver in those cases, the messages are not being misused in those states, relative to the population.

Under my proposed definition, however, the sender type that is uniquely identifying each state of the world by some signal – type S_1 – behaves deceptively in every state of the world. In each state of the world, there is only one message that is sent by S_1. According to the population, though, S_1 is misusing whichever message is sent in that state.[3] Furthermore, the use of each message is actually deceptive in its respective state because the use is beneficial to the sender and detrimental to the receiver.

[3]Straightforward calculations give, for instance, $J(m_1, q_1; S_1, \sigma_P) = \log(3/2) > 0$ and $J(m_1, q_2; S_1, \sigma_P) = -\infty < 0$. Similar results hold for the other state-message combinations that type S_1 employs.

For instance, in state q_1, senders of type S_1 send message m_1. Since the receiver does not know that the sender is actually type S_1, she has to respond as if the message was sent by a random member of the sender population. The best response in that case would be to take action a_3, netting the sender and receiver each a payoff of 10. If the receiver knew, however, that the actual state of the world was q_1, she would prefer to do action a_1, earning the sender a payoff of only 8 and the receiver a better payoff of 12.

In addition to exhibiting nearly universal deception, this example also illustrates another important possibility: under my definition, it is possible to be deceptive by trying to tell the whole truth.[4] This identification is not possible under Skyrms's definition. A message cannot contain misinformation when it is trying to tell the whole truth, and without misinformation, there is no deception.

Skyrms could say that the population is using each of the messages deceptively. The use of each message in the population average σ_P does contain misinformation to the benefit of each sender and detriment of each receiver. This is not enough. Populations are not the sorts of things that we usually want to say are deceptive; individuals (or types of individuals) are.

It might be tempting to look at this example and think that type S_1 is clearly not being deceptive. It is, after all, trying to tell the whole truth. This would be mistaken. To see why, imagine a similar situation where the agents were fully rational and the senders of type S_1 were far rarer.[5] Based on the structure of the game, a receiver could easily infer that no sender should reveal the entire truth. For, there is a more rational strategy: using one

[4] Although this might not be the most common form of deception, it is still possible. Sutter, in the context of intentional deception, claims, "telling the truth should be classified as intended deception if the sender chooses the true message with the expectation that the receiver will not follow the sender's (true) message." (2009, pp. 47–8) He goes on to argue that this kind of deception is, in fact, more common than one might initially expect. This line of reasoning is quite similar to the argument I gave for why uniform behavior is not deceptive. The key there was that expectations were not violated, for some appropriate sense of expectation.

[5] For signalling games in an evolutionary context, agents are usually considered to be automatons. I use the term "rational" here to distinguish this situation from that sort of context.

message as a synonym for two states of the world. If these receivers had been interacting with a population of senders replete with types S_2 and S_3, but came across an S_1 sender instead, the receiver would rationally expect message m_1, for example, to be used in the way the population overall uses it. For, the receiver can't observe the type of the sender.

Despite the fact that senders of type S_1 would be trying to tell the truth, they are using messages differently than the rest of the population in a way that is deceptive. That is what is responsible for the deception. Although that story includes the supposition that truth-tellers are rare, even if their proportion is increased, similar conclusions should be reached.

3.3.3 Strongly Universal Deception

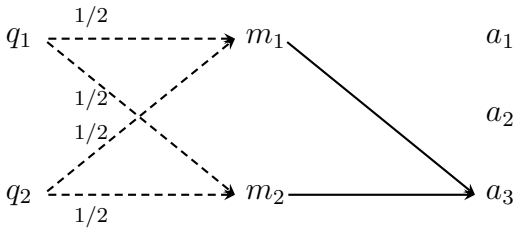
Although the previous example demonstrated that nearly universal deception is possible in equilibrium under my definition, that was not the final goal. A slightly simpler game can exhibit the strongest form of universal deception I have described above – the same one Skyrms's example was supposed to illustrate. This details of this signalling game are presented in Figure 3.4. In many ways, the game is quite similar to considering only half of the previous example. There are two states of the world. Each receiver would most like to uniquely identify the state, but the senders would prefer that the receiver did action a_3 instead of uniquely identifying whether the world was in state q_1 or q_2.

One possible equilibrium of this game is given in Figure 3.4b. Senders randomize which message is sent in each state of the world. Given that the states of the world are equally likely given any message, the receivers' best response is to perform action a_3, as a guaranteed payoff of 6 is superior to a 50% chance of 10 and a 50% chance of 0.

Again, to apply my definition of deception to this situation, we need to know how the

	a_1	a_2	a_3
q_1	$5, 10$	$0, 0$	$6, 6$
q_2	$0, 0$	$5, 10$	$6, 6$

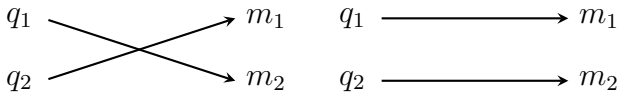
(a) State-Act Payoffs



(b) Equilibrium Behavior

Figure 3.4: Strongly Universal Deception Example

population of senders is composed. One possible composition is an even mixture of the sender types S_1 and S_2, which are pictured in Figure 3.5.[6] Each of these senders tries to tell the entire truth, but they create interference with each other in the population at large.



(a) Type S_1 (b) Type S_2

Figure 3.5: Strongly Universal Deception Sender Types

Both types of sender in this case are deceptive in every state of the world. For instance, a sender of type S_1 deceptively uses message m_2 in the state of the world q_1. To see this, first note that the message is misused in that state: $J(m_2, q_1; S_1, \sigma_P) = 1$ and $J(m_2, q_2; S_1, \sigma_P) = -\infty$. This misuse is also deceptive. In an actual interaction, the sender and receiver would each receive a payoff of 6 after the receiver performed action a_3. If the receiver knew the actual state of the world, however, she would perform action a_1 instead, increasing her payoff to 10 and lowering the sender's payoff to 5.

Given the symmetries of the equilibrium, it should be straightforward to see that analogous calculations hold for both sender types and for both states of the world. This example, therefore, is the kind we were trying to construct: an equilibrium exhibiting strongly universal deception in the same sense that Skyrms's example did for his definition.

[6]The labelling of these strategies might appear strange. The reason for this order is the encoding of strategies for the simulations performed in the next section.

3.4 Evolutionary Significance

The previous example shows that universal deception is a stable possible outcome in an evolutionary context. However, it does not indicate how likely such a situation is to actually arise. The answer to that question will depend on the dynamics used and the stability properties of the equilibrium under those dynamics. It could very well be that all examples of universal deception are unstable rest points of the dynamics, making their emergence from an evolutionary process essentially impossible. Or, they could not even be rest points at all. It will turn out that the equilibrium I've described is not as rare as those possibilities. It is, however, not likely to arise naturally.

3.4.1 Replicator Dynamics

My investigation of this question will make use of the one-population replicator dynamics. The continuous time replicator dynamics will be used to discuss the abstract theoretical properties of universal deception. Simulations will then be presented using the discrete-time replicator dynamics to estimate how likely such an equilibrium is to arise.

Both the continuous-time and discrete-time replicator dynamics govern the evolution of a population based on the fitness of each type of individual in an infinite population. If the type's fitness is greater than the average fitness of the population, its representation will increase proportionally. If, on the other hand, its fitness is less than the average fitness of the population, its representation will decrease proportionally. In a standard game, the fitness of an individual or type is identified with its payoff.

Mathematically, the continuous-time dynamics are given by a system of differential equations

(Weibull, 1995, p. 72):

$$\dot{x}_i = \left[u(e^i, x) - u(x, x)\right] x_i \tag{3.1}$$

Here, x is the vector of population proportions, and x_i represents the proportion of type i in the population. The notation e^i is the pure strategy played by type i, and u is the payoff function for the game being played. Writing $u(e^i, x)$ indicates the payoff that type i earns when playing against the population at large, and $u(x, x)$ indicates the average payoff for all types in the population.

The discrete-time dynamics are similar. It uses finite time-steps to update the population proportions instead of a continuous differential equation (*ibid.*, p. 123):

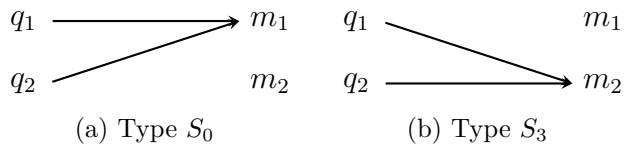
$$x_i(t+1) = \frac{\alpha + u[e^i, x(t)]}{\alpha + u[x(t), x(t)]} x_i(t) \tag{3.2}$$

In this system of equations, $t = 0, 1, 2, \ldots$ is the discrete time step, $x(t)$ is again the vector of population proportions for each type at time t, $x_i(t)$ is the proportion of type i at time t, and u is the payoff function for the game. The parameter $\alpha \geq 0$ is the background fitness for each type.

3.4.2 Stability of the Equilibrium

The equilibrium demonstrating strongly universal deception (Figure 3.4) is part of a manifold, $\mathcal{M}$, of Nash equilibria. Once receivers are responding to all messages by taking action a_3, there are many sender strategies that perform equally well. Switching a small segment of a sender population comprised of types S_1 and S_2 (Figure 3.5) to instead play type S_0 or S_3 (Figure 3.6) results in another equilibrium, as long as the ratios $\frac{Pr(q_1|m_1)}{Pr(q_2|m_1)}$ and $\frac{Pr(q_1|m_2)}{Pr(q_2|m_2)}$ each fall between $\frac{2}{3}$ and $\frac{3}{2}$. The reason for that within that range, performing action a_3 is

the best response to any message for a receiver. Each of these equilibria also exhibits deceptive behavior, though not universal deception in the strongest sense. Exactly how much deception is found depends on the exact composition of the population.



(a) Type S_0 (b) Type S_3

Figure 3.6: Alternative Universal Deception Sender Types

The equilibria on the interior of the manifold $\mathcal{M}$ (but not necessarily those on the boundary) are neutrally stable. This means that for any equilibrium on the interior of the manifold, no nearby strategy resulting from a small perturbation is a better response to the perturbed population than the equilibrium response would be.[7] Under the replicator dynamics, what this means is that small perturbations of a population on the interior of the manifold do not result in a trajectory that leaves the vicinity of the manifold.

Part of this is easy to understand: perturbations that move along the manifold itself result in equilibria, and all of the equilibria on the manifold are equally good responses to each other. Determining that any interior point on the manifold is at least as good of a response to any perturbation to a population off of the manifold as the perturbed population is to itself is slightly more complicated.

First note that small perturbations away from $\mathcal{M}$ result in some proportion of the population playing a new sender strategy, a new receiver strategy, or both.[8] If the perturbation introduces only new sender strategies, then the perturbation is on the manifold of equilibria. If the perturbation introduces only new receiver strategies, then the receiver strategy of al-

[7]The precise condition is that a population $x \in \Delta$ is neutrally stable when for every strategy $y \in \Delta$ there is some $\bar{\epsilon}_y \in (0,1)$ such that the inequality $u[x, \epsilon y + (1-\epsilon)x] \geq u[y, \epsilon y + (1-\epsilon)x]$ holds for all $\epsilon \in (0, \bar{\epsilon}_y)$. (Weibull, 1995, p. 46) Here Δ is the space of possible populations playing the game, or equivalently the space of possible mixed strategies, and u is the payoff function for the game.

[8]Recall that the dynamics that will be used are one-population. To accommodate this, the game is symmetrized so each individual has both a sender and a receiver strategy. The type of an individual is defined as the combination of those.

ways choosing a_3 is superior. So, the interesting case is when a perturbation simultaneously introduces new sender and receiver strategies.

Suppose, then that x^* is some equilibrium in the interior of $\mathcal{M}$. So, the population at x^* is behaving as in Figure 3.7, where $\frac{2}{3} < \frac{\alpha}{\beta} < \frac{3}{2}$ and $\frac{2}{3} < \frac{1-\alpha}{1-\beta} < \frac{3}{2}$.

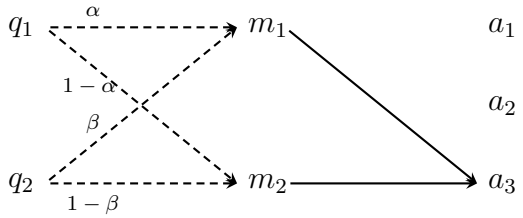


Figure 3.7: An Arbitrary Equilibrium on the Manifold $\mathcal{M}$

Consider then a perturbation where a small proportion $\epsilon > 0$ of the population changes receiver strategy and a small proportion $\delta > 0$ of the population changes sender strategy. Are either of these new strategies better against the perturbed population than the originals would be?

First, consider the receiver strategy component. For δ sufficiently small, the average sender behavior in the perturbed population still falls within the range where the receiver strategy to always take action a_3 is a strict best response. So, no alterations to the receiver strategy will produce an advantage.

Now, consider the sender strategy component. Against the proportion of the population that was not perturbed, the sender strategy is just as fit as the established senders. For, all un-perturbed receivers take action a_3 given any message. So, any difference that part of the perturbation makes can only come from the interaction with the new receivers. So, could any receiver strategy be introduced that a sender would prefer to interact with? The answer is no. The receiver taking action a_3 in any state of the world is the best possible outcome for a sender.

So, the types from the established population do at least as well against the perturbed

population than any mutants do. Thus, the points on the interior of $\mathcal{M}$ are neutrally stable.

3.4.3 Simulation Results

The equilibrium that demonstrates the possibility of strongly universal deception is neutrally stable. But, what does that mean for its evolutionary significance? For one, it tells us that the equilibrium exhibiting strongly universal deception is Lyapunov stable under the continuous-time replicator dynamics. (Weibull, 1995, p. 104) This means that trajectories nearby the equilibrium stay nearby the equilibrium. What this does not tell us, however, is how likely an population is to reach a point at least nearby the equilibrium. To generate an estimate of that propensity, I turn to simulation.

Simulations were run on the symmetrized form of the game using the discrete-time replicator dynamics. The results are summarized in Table 3.1.[9] In general, this table shows that the manifold of equilibria $\mathcal{M}$ has a very small basin of attraction. Furthermore, arriving exactly at the equilibrium described in Figure 3.4b is exceedingly rare (it was never seen in simulation). Some simulations did come close, however. Of those simulations that ended with a population on $\mathcal{M}$, a large proportion of interactions were deceptive (60–70%), and in

[9] In the table, the column labels have the following meanings:

Runs The number of times the simulation was performed with those parameter settings.

α The value of the α parameter in the discrete-time replicator dynamics equations.

Mixed The number of simulations that ended with a polymorphic population.

On $\mathcal{M}$ The number of simulations that ended with a polymorphic population on the manifold $\mathcal{M}$.

Dec. The number of simulations exhibiting deceptive behavior by some sender against some receiver.

Dec. on $\mathcal{M}$ The number of simulations exhibiting deceptive behavior by some sender against some receiver on the manifold $\mathcal{M}$.

Mean % Of those simulations that exhibited deception, the average percentage of interactions that were deceptive.

Max % Of those simulations that exhibited deception, the maximum percentage of interactions that were deceptive.

one population, approximately 92% of interactions were deceptive.

Runs	α	Mixed	On $\mathcal{M}$	Dec.	Dec. on $\mathcal{M}$	Mean %	Max %
5000	0.0	614	4	4	4	72%	92%
5000	0.1	595	2	2	2	71%	74%
5000	1.0	584	4	4	4	61%	75%

Table 3.1: Simulation Results for the Game in Figure 3.4a

One major factor that might have affected these results is that the signalling system equilibria Pareto dominate those on the manifold in the symmetrized form of the game. Adjusting the payoffs of the game to make that untrue, mixed equilibria in general become far more common results of the dynamics. Another set of simulations was run using the payoffs in Figure 3.8.

	a_1	a_2	a_3
q_1	$1, 10$	$0, 0$	$6, 6$
q_2	$0, 0$	$1, 10$	$6, 6$

Figure 3.8: Payoff Scheme 2 for Simulations Regarding Universal Deception

With these payoffs, the manifold $\mathcal{M}$ is still neutrally stable.[10] But, the signalling system equilibria no longer Pareto dominate the receiver always taking action a_3. Given these payoffs, simulations converged, relatively speaking, much more frequently to $\mathcal{M}$, as can be seen in Table 3.2.[11] Universal deception itself was not observed as a direct result, and in global terms, the basin of attraction of the manifold is still quite small.

Similarly to the previous results, however, the number of deceptive interactions on average was quite high. The mean percentages across all final populations were above 70% and in each of the treatments at least one population ended with greater than 90% of interactions being deceptive.

These results are also fragile to payoff perturbations. Changing the payoffs again to those

[10]The edges of the manifold of equilibria do not shift, as the receiver payoffs are the same as in the previous example.

[11]The table key is identical to the one given in footnote 9.

Runs	α	Mixed	On $\mathcal{M}$	Dec.	Dec. on $\mathcal{M}$	Mean %	Max %
5000	0.0	4123	32	32	32	71%	94%
5000	0.1	4157	28	28	28	76%	95%
5000	1.0	4055	41	41	41	71%	96%

Table 3.2: Simulation Results for the Game in Figure 3.8

reflected in Figure 3.9, polymorphic equilibria persist, but deception nearly disappears. The reason for this might be that in the symmetrized game, in state q_1, coordinating with the receiver to take action a_1 and taking action a_3 are equivalent. So, the potential advantage of always performing action a_3 is reduced, enabling signalling systems to be more prevalent.

	a_1	a_2	a_3
q_1	$1, 10$	$0, 0$	$6, 5$
q_2	$0, 0$	$1, 10$	$6, 6$

Figure 3.9: Payoff Scheme 3 for Simulations Regarding Universal Deception

Runs	α	Mixed	On $\mathcal{M}$	Dec.	Dec. on $\mathcal{M}$	Mean %	Max %
5000	0.0	2856	1	1	1	71%	71%
5000	0.1	2909	0	0	0	N/A	N/A
5000	1.0	2847	0	1	0	8%	8%

Table 3.3: Simulation Results for the Game in Figure 3.9

3.4.4 Discussion

So what can we say about the possibility of universal deception? On the one hand, Skyrms is correct: strongly universal deception is possible in equilibrium, as we saw in the game from Figure 3.4. So, Kant's claims, read at face value appear to be incorrect.

On the other hand, however, Kant appears to be correct in some sense. One of the main themes in the situations Kant describes is that attempting to establish a situation that is universally deceptive is unstable and will cause the system to collapse. Although this is not

always the case in the signalling games I've considered, it is approximately correct most of the time.

The basins of attraction for the manifold of equilibria on which the example of strongly universal deception lies is quite small. In a large majority of cases, if the sender population begins near to a population that could possibly be universally deceptive, the receiver component evolves to not be deceived and induces the sender population to change in the process. Very often this results in a signalling system equilibrium being established. In some sense, then, Kant appears to have been largely correct. A population of senders all trying to be deceptive undermined the system.[12]

Changing the payoff structure of the game altered the results substantially, but not very much in favor of the evolution of universal deception. More simulations resulted in polymorphic populations, but few of those were on the equilibrium manifold $\mathcal{M}$. Of those that did end up on $\mathcal{M}$, a few were close to exhibiting universal deception, but none showed fully universal deception. In this sense, Kant seems to have been correct again. The result was often not a signalling system equilibrium, but it was quite rare to see a result even close to universal deception.

So, it appears that Kant was right to a certain extent. Although it is logically and evolutionarily consistent that a population should exhibit universal deception, the likelihood of this developing naturally is quite small. Attempts to establish such a system *de novo* would likely undermine themselves, resulting in other, non-deceptive kinds of behavior instead.

The results above obviously don't apply to all possible signalling games that might be constructed. However, I have tried to show that strongly universal deception is both rare and fragile.

[12] In the simulations, there is obviously no intention, or "trying" to be deceptive. This phrasing is included to make it understandable in terms closer to those used by Kant.

Chapter 4

Self-Deception

4.1 Why Self-Deception?

Humans and other animals regularly deceive other organisms. Detailed reasons for this can vary, but in the end there is clearly some benefit to be gained by deception in these situations. Not only do humans and other animals deceive conspecifics (and in some cases even non-conspecifics – see (Searcy and Nowicki, 2005)), they also appear to deceive themselves. This self-deception is more of a puzzle than other-deception. Is deceiving oneself even possible? If it is, what benefit could there be?

Davidson, for example, answers the first question in the negative.[1] He notes, "We can now see the difficulty in taking the notion of lying to oneself too literally: it would require that one perform an act with the intention both that that intention be recognized (by oneself) and not recognized (since to recognize it would defeat its purpose). We had better, then, take the expression 'lying to oneself' as a kind of metaphor. . . ." (Davidson, 1998, p. 3) Fingarette, among others, has tried to avoid this problem by relying on the notion of unconscious systems

[1] Davidson's full story is more nuanced than the simple quote that follows, but the quote illustrates one of the main problems that others have also noted with the concept of self-deception.

that can work independently of the consciousness. He writes, "In general, the person in self-deception is a person of whom it is a patent characteristic that even when normally appropriate he *persistently* avoids spelling-out some feature of his engagement in the world." (Fingarette, 1969, p. 46) The notion of spelling-out is the presentation of information to the consciousness in a certain way:

> To spell-out, I have said, is to be explicitly aware of; it is to pay conscious attention to. We might speak of this as the 'strongest' sense of 'conscious'. By contrast, there is the 'weaker' sense of 'conscious' in: 'Though struck a heavy blow, he remained fully conscious'; or 'he lost consciousness'. Also contrasting with selling-out is another 'weaker' sense of 'conscious' in: 'Are you conscious that you are shuffling the cards?' 'Yes, of course, I'm perfectly conscious that I'm doing it, but that hasn't distracted me from what you are saying; I'm paying attention only to what you say and nothing else enters my mind.' (*ibid.*, p. 44)

Fingarette's suggestion seems generally applicable, even to non-human animals. Although we cannot be sure that non-human animals have consciousness at all, let along one like our own, it seems possible enough to warrant further investigation on these grounds. So, I will proceed with a story similar to Fingarette's in mind: self-deception is possible, and it is based on what is spelled-out to the consciousness – what is explicitly presented.

The second question is perhaps more interesting and one on which some progress can be made. Assuming self-deception is possible, why have animals evolved a capacity to engage in it? Prima facie, it seems that there could be no evolutionary advantage. We rely on veridical representations of the external world to successfully navigate, acquire food, and not fall off cliffs. On the other hand, though, there are definitely reasons that humans might engage in self-deception: enhancing self-confidence, for example. Though these benefits definitely make us feel better, it is unclear that they would have been evolutionarily advantageous. In

animals, the picture is even fuzzier.

Robert Trivers and others (Trivers, 2000, 2011; von Hippel and Trivers, 2011) have suggested that self-deception evolved to assist in deceiving others. The general idea is that attempting to deceive carries with it involuntary cues that a deception is occurring. Examples of this might be blushing or unusually terse replies due to increased cognitive load. Receivers paying attention to these cues could then avoid being deceived.

Self-deception's role would be to avoid giving those cues. By deceiving oneself that the state of the world was actually something advantageous (instead of something that would be beneficial to be deceptive about), reporting on that state would then be much more similar to telling the truth, from a conscious point of view. There wouldn't be additional cognitive load, since you aren't trying to suppress the fact that you are being deceptive, and cues like blushing would be less likely to trigger.

Using the tools of evolutionary game theory and formal models of signalling, a clearer answer can be provided, at least in terms of the possibility of Trivers et al.'s theory. Using a three-player signalling model and including the possibility of a receiver detecting cues of deception, we can see how likely self-deception is to develop and how much other-deception is present.

4.2 The Model

The model I will employ to investigate the plausibility of Trivers's claims is a three-player signalling game. The three players are Sender 1, Sender 2, and the Receiver. They are supposed to correspond to the Unconscious Mind of a agent, the Conscious Mind of the same agent, and an external Receiver.

The world can be in three possible states, and the sending agent and external receiver can

be in one of two Relationships: common interest or partial common interest.

The three players interact by a chain of signals. First, the Unconscious Mind observes the state of the world, corresponding to raw sensory input. The Unconscious Mind then chooses a Representation to present to the Conscious Mind from a set naturally corresponding to the actual states.[2]

The Conscious Mind observes the Representation and which Relationship she stands in to the Receiver. She chooses a signal to send. There are exactly as many signals available as there are states of the world, so the Conscious Mind could possibly uniquely identify the correct state.

The Receiver observes the signal and the Relationship with the sending agent and chooses an action to take. There is one additional action available to the Receiver than the usual suite corresponding to the states of the world. This extra action is to Investigate further. If the Receiver chooses to Investigate, with some exogenous probability, she correctly identifies the representation that was presented to the Conscious Mind and best-responds to that. If she Investigates and incorrectly identifies the representation, she thinks a randomly selected representation was presented instead.

The Investigate act is supposed to correspond to the ability to pay close attention to uncontrollable signs of deception, such as blushing, cognitive load, and nervousness. (see, e.g. Trivers, 2000, 2011; von Hippel and Trivers, 2011) Choosing to Investigate also carries an exogenously specified cost to the Receiver that does not depend on the success or failure of the investigation.

Once an action is selected, all three players receive a payoff according to the Relationship for that interaction. The payoff for the Unconscious and Conscious players is modified by

[2]This natural correspondence will be identified by the representations and states of the world being written identically. For purposes of the simulations performed, however, the symbols were distinct.

a parameter that gives an incentive for accurate representations. This parameter indicates the proportion of fitness derived from interaction with the receiver, as opposed to decision problems faced by the organism containing the Unconscious and Conscious senders where accurate representation would be beneficial.

Formally, the model is a three-player signalling game with states of the world $Q = \{q_1, q_2, q_3\}$, relationships $S = \{s_C, s_P\}$, messages $M = \{m_1, m_2, m_3\}$, actions $A = \{a_1, a_2, a_3, a_I\}$, Sender 1 strategy $U\colon Q \to Q$, Sender 2 strategy $C\colon Q \times S \to M$, and Receiver strategy $R\colon M \times S \to A$. The probability of success in trying to Investigate is $p \in [0, 1]$, and the cost of choosing to Investigate is c. The parameter representing the proportion of sender fitness derived from external interactions is $\varepsilon \in [0, 1]$.

The behavior of agents in a particular relationship (s_C or s_P) can be represented in a traditional behavior map. For example, the behavior map in Figure 4.1 shows the Unconscious pooling states q_2 and q_3 on to the same representation, and the Receiver choosing to Investigate if she receives the signal m_3 but not otherwise.

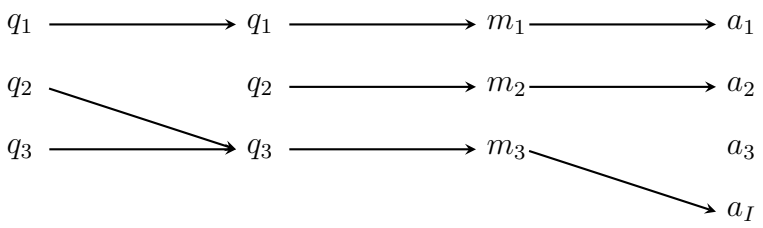


Figure 4.1: Example Behavior in the Self-Deception Model

The basic state-act payoffs for the common interest and partial common interest relationships are given in Figures 4.2a and 4.2b respectively. The term $\varrho(q_i, U) = (1 - \varepsilon)\eta(q_i, U(q_i))$ represents the proportion of the Unconscious and Conscious payoffs derived from decision problems instead of external interactions. The function $\eta(\cdot, \cdot)$ is equal to 1 if the two arguments are equal and 0 otherwise. The list of payoffs are first for the Unconscious sender, then the Conscious sender, and finally for the Receiver. In a particular interaction, the Receiver's payoff might also be affected by the cost of investigation, but this is not shown in the table.

	a_1	a_2	a_3
q_1	$\varepsilon + \varrho(q_1, U), \varepsilon + \varrho(q_1, U), 1$	$\varrho(q_1, U), \varrho(q_1, U), 0$	$\varrho(q_1, U), \varrho(q_1, U), 0$
q_2	$\varrho(q_2, U), \varrho(q_2, U), 0$	$\varepsilon + \varrho(q_2, U), \varepsilon + \varrho(q_2, U), 1$	$\varrho(q_2, U), \varrho(q_2, U), 0$
q_3	$\varrho(q_3, U), \varrho(q_3, U), 0$	$\varrho(q_3, U), \varrho(q_3, U), 0$	$\varepsilon + \varrho(q_3, U), \varepsilon + \varrho(q_3, U), 1$

(a) Full Common Interest

	a_1	a_2	a_3
q_1	$\varrho(q_1, U), \varrho(q_1, U), 1$	$\varepsilon + \varrho(q_1, U), \varepsilon + \varrho(q_1, U), 0$	$\varrho(q_1, U), \varrho(q_1, U), 0$
q_2	$\varepsilon + \varrho(q_2, U), \varepsilon + \varrho(q_2, U), 0$	$\varrho(q_2, U), \varrho(q_2, U), 1$	$\varrho(q_2, U), \varrho(q_2, U), 0$
q_3	$\varrho(q_3, U), \varrho(q_3, U), 0$	$\varrho(q_3, U), \varrho(q_3, U), 0$	$\varepsilon + \varrho(q_3, U), \varepsilon + \varrho(q_3, U), 1$

(b) Partial Common Interest

Figure 4.2: State-Act Payoffs for the Self-Deception Game

4.3 Self-Deception and Other-Deception

Defining other-deception in this model is a straightforward application of the definition I constructed in chapter 2 as long as we can construct appropriate populations for comparison. Misuse of a message is determined according to the representation that the Conscious sender was presented. Sender benefit and receiver detriment, though, are calculated according to the actual state of the world.

Trying to pick out self-deception is more problematic, however. In the payoff tables from Figure 4.2, the Unconscious and Conscious senders have pure common interest. There could never be a benefit for the Unconscious sender or a detriment for the Conscious sender in any possible behavior. Without the possibility of benefits and detriments, there could be no deception.

This problem is resolvable. The payoffs listed are those relevant to evolutionary fitness. That is not what we usually think about in the context of self-deception. Instead, the conscious mind is thought to be deceived when it does not have a veridical representation of reality.

The solution then is to construct an alternate set of payoffs for the Conscious to use in identifying self-deception. The Unconscious cares not about veridical representation of the world. It only "cares" about evolutionary fitness. So, the standard fitness payoffs (as determined by the actions of all three players) are still appropriate to use in that case.

The other issue in constructing these alternative payoffs is that the Conscious doesn't take an action in the standard sense. In fact, whether the representation given is veridical is only determined by the choice of the Unconscious. So, there is no response officially taken by the Conscious, let alone a best-response.

To construct the self-deception situation, then, we must also create placeholder actions of the understanding for the Conscious. There is no choice involved in this case. The response is pre-programmed.

This leaves us with a smaller companion game to the full model described above. The Unconscious in this companion game behaves just as in the full model. The Conscious, though, is programmed to take a pseudo-action from among α_1, α_2, and α_3, according to the diagram in Figure 4.3b. In that figure, the q_i are the representations provided by the Unconscious, not the actual states of the world. The interactions result in the state-act payoffs in Figure 4.3a. The function $u_U(q_i, a)$ is the payoff to the Unconscious in the full game based on all three players' actions.

	α_1	α_2	α_3
q_1	$u_U(q_1, a), 1$	$u_U(q_1, a), 0$	$u_U(q_1, a), 0$
q_2	$u_U(q_2, a), 0$	$u_U(q_2, a), 1$	$u_U(q_2, a), 0$
q_3	$u_U(q_3, a), 0$	$u_U(q_3, a), 0$	$u_U(q_3, a), 1$

(a) State-Act Payoffs

$q_1 \longrightarrow \alpha_1$

$q_2 \longrightarrow \alpha_2$

$q_3 \longrightarrow \alpha_3$

(b) Conscious Behavior

Figure 4.3: Companion Self-Deception Game

4.4 Simulations

Simulations were run using Herrnstein-Roth-Erev reinforcement learning dynamics (Herrnstein, 1970; Roth and Erev, 1995). Under these dynamics, the three players reinforce chosen behaviors based on the success that they had in using them. Choosing a behavior in the future is then proportional to their past success. All initial weights were equal and set to 1.

Simulations were performed for values of p between 0 and 1 at intervals of 0.1 and for values of c between 0 and 0.6 also at intervals of 0.1. The parameter ε representing the proportion of the Unconscious and Conscious fitness derived from external interactions was also varied, taking on values of 0.75 (75% of the fitness comes from external interactions), 0.9 (90% of the fitness comes from external interactions), and 1.0 (all of the fitness comes from external interactions).[3] Each parameter setting was run 1000 times.

4.4.1 Results for Epsilon at 0.75

Setting $\varepsilon = 0.75$, even with a fairly high proportion (25%) of the Unconscious and Conscious fitness determined by a decision problem based on the internal representation of the state of the world, a variety of other-deceptive and self-deceptive behavior was readily observed.

Representations of the state of the world were largely veridical. Of the 1000 duplicate runs at each parameter setting, an average of 904 of them resulted in the Unconscious truthfully representing the state of the world with probability greater than 99.5%. The more costly inspection was, the more veridical the representations were likely to be as well (Figure 4.4).

Veridical representations in this case were likely driven both by the proportion of fitness garnered outside of interaction with the Receiver and by the likelihood that the Receiver

[3]The $\varepsilon = 0.9$ setting did not explore the cost parameter as much (ranging only from 0 to 0.4). A few additional parameter settings were explored, but the detailed results were not radically different from similar parameter settings that were explored more fully.

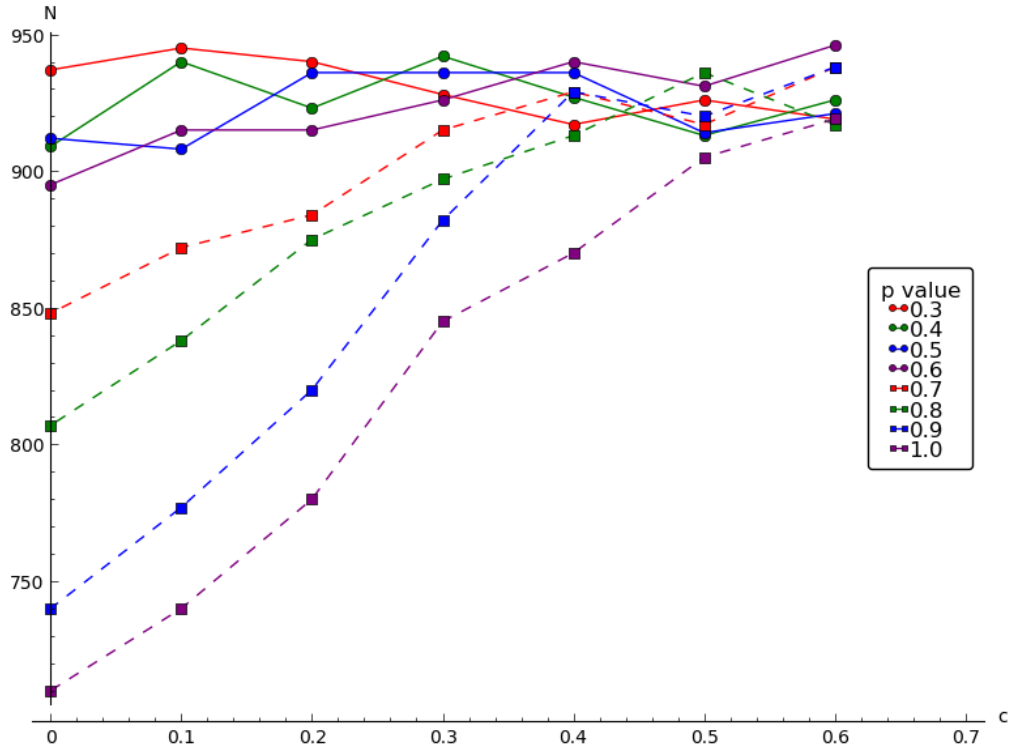
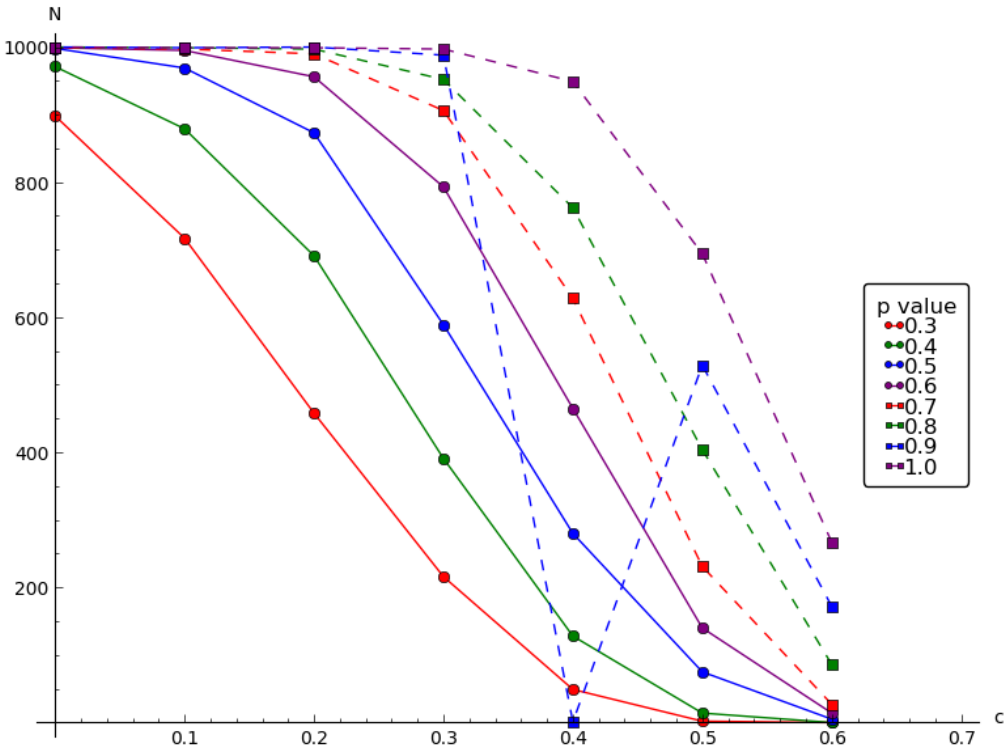


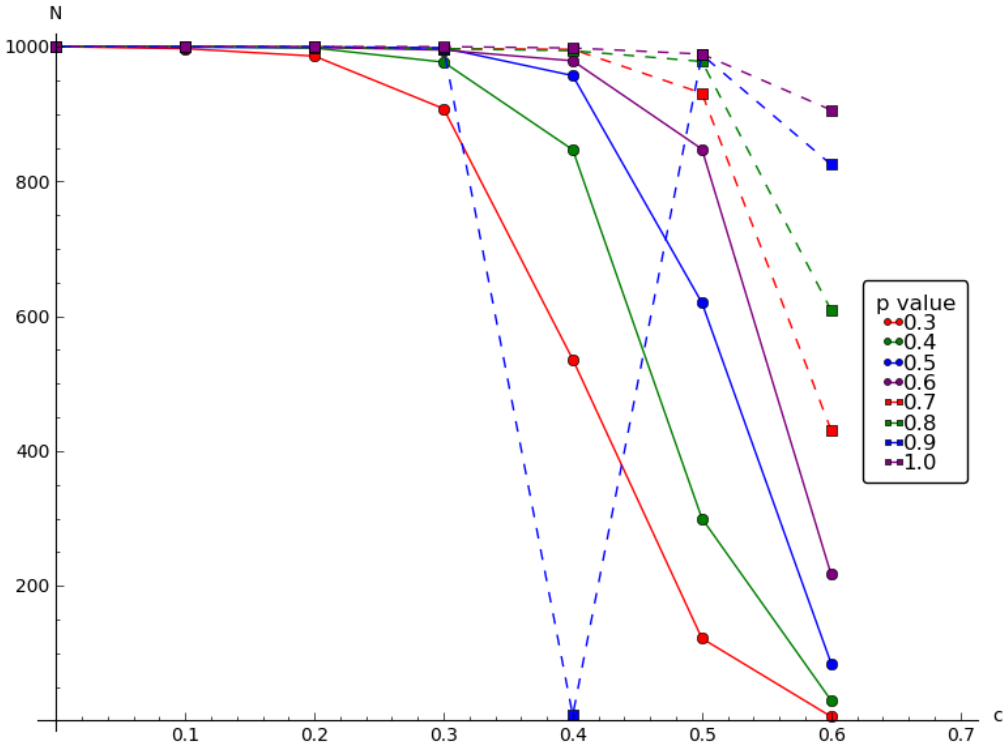
Figure 4.4: Number of Runs with Virtually Totally Veridical Representation, $\varepsilon = 0.75$

would choose the Inspect action. For low inspection costs, the Receiver nearly always chose to use the Inspect action at least 1% of the time (Figure 4.5). This was the case both when there was full common interest (s_C) and when there was only partial common interest (s_P). In fact, the Receivers regularly chose to Inspect with fairly substantial frequency, especially when the probability of success was high (Figure 4.6).

The veridical nature of the representations also corresponded with a low incidence of self-deception (Figure 4.7). When the Conscious and Receiver had perfect common interest (s_C), of the 1000 duplications at each parameter setting, an average of 62 showed some self-deceptive behavior. With only partial common interest (s_P), the average number of runs with self-deceptive behavior increased, but only to 82. When the probability of correct inspection was sufficiently high ($p \geq 0.7$), there was also a noticeable trend of the incidence of self-deception decreasing as the cost of the inspection action increased.

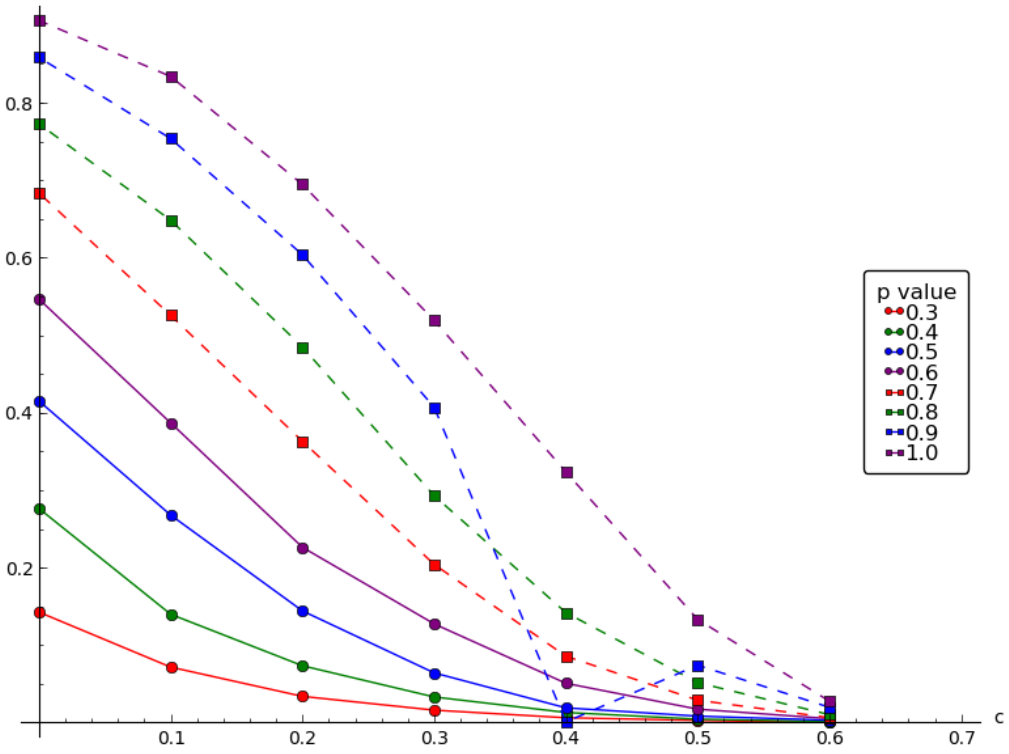


(a) Full Common Interest

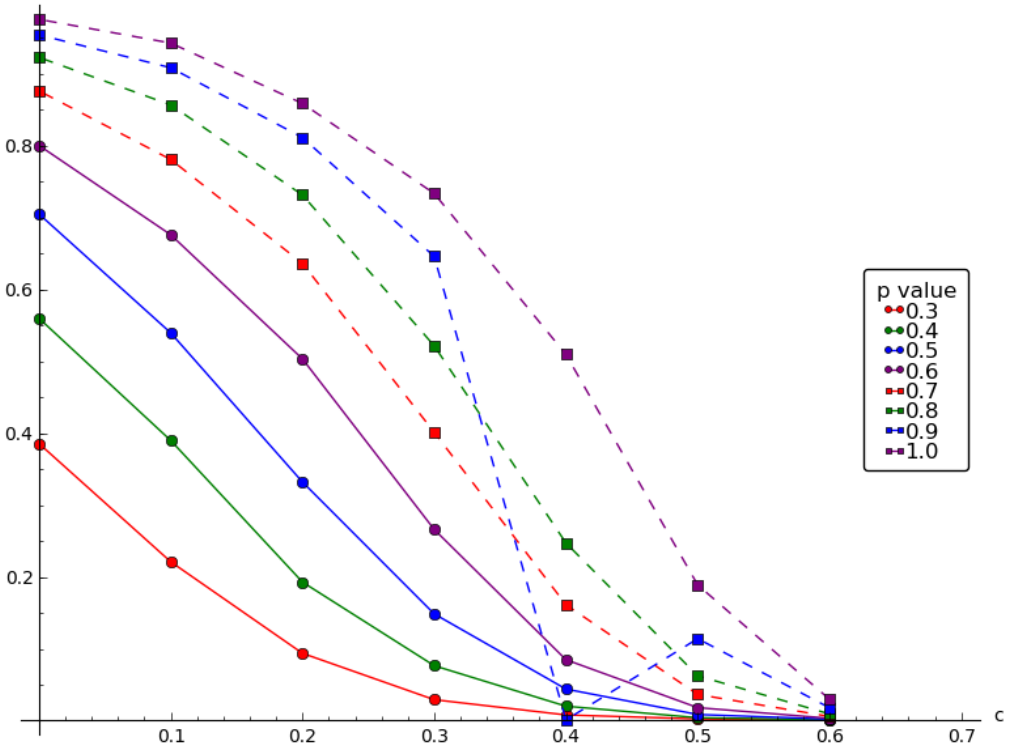


(b) Partial Common Interest

Figure 4.5: Number of Runs with Inspect Use, $\varepsilon = 0.75$

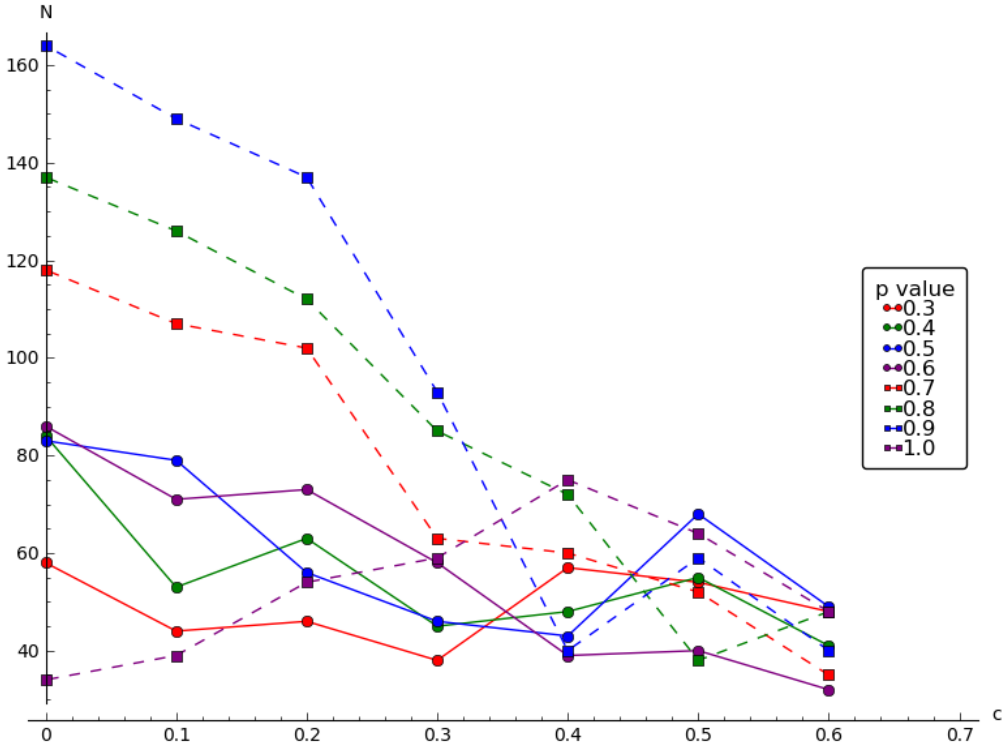


(a) Full Common Interest

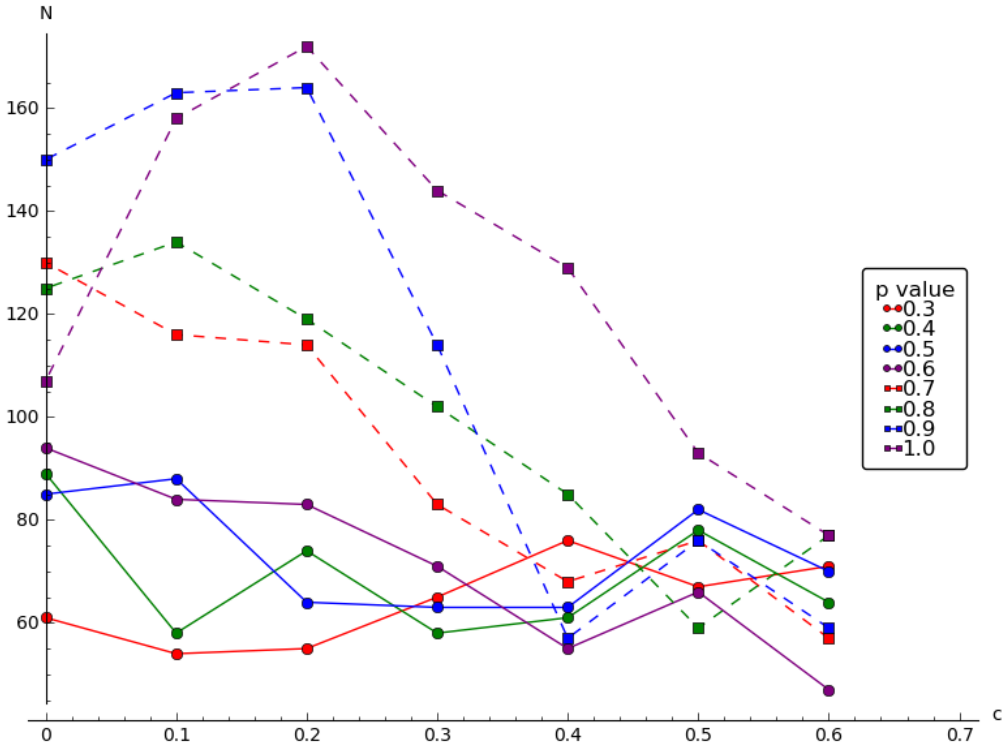


(b) Partial Common Interest

Figure 4.6: Average Percentage of Inspect Use, $\varepsilon = 0.75$



(a) Full Common Interest



(b) Partial Common Interest

Figure 4.7: Number of Runs with Self-Deceptive Behavior, $\varepsilon = 0.75$

Conscious deception, however, was rampant in the case of partial common interest (Figure 4.8a).[4] Conscious deception, in this context, means that the Conscious sender misuses a message relative to the representation it was provided in a way that beneficial the Conscious and detrimental to the Receiver. A substantial portion of those conscious deceptions occurred when the Receiver chose the Inspect action and the inspection revealed the actual representation (Figure 4.8b).

It is also possible to consider deception by the sender organism as a whole. To do this, we construct a sender behavior that skips the representation step.[5] This whole-organism behavior is, by necessity, even more prevalent than the conscious deception (Figure 4.9a). With these parameter settings, a small but substantial number of runs had whole-organism deception that was the result of both self-deception and conscious deception (Figure 4.9b).[6] This makes sense given the relatively low incidence of self-deception by itself. Some runs also featured whole-organism deception that was the result of *only* self-deception, but the average number of such runs was only 10 in these settings.

Considering this whole-organism behavior also reveals an interesting phenomenon. There can be deceptive behavior by the whole organism that is neither consciously deceptive nor self-deceptive. Instead, it appears to emerge from the combination of the Unconscious and Conscious behavior. The phenomenon is rare for the parameter values that were explored, but might be more prevalent in other ranges (Figure 4.9c).

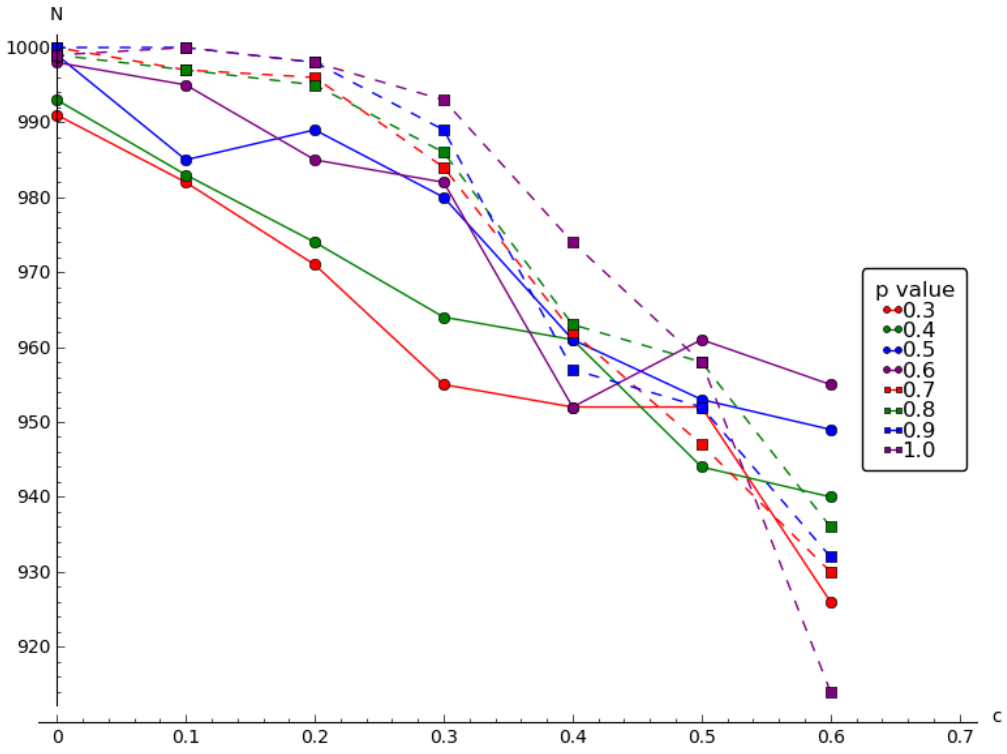
4.4.2 Results for Epsilon at 0.9

Reducing the proportion of the Unconscious and Conscious fitnesses that are determined independently of their interaction with the Receiver shows some more interesting results

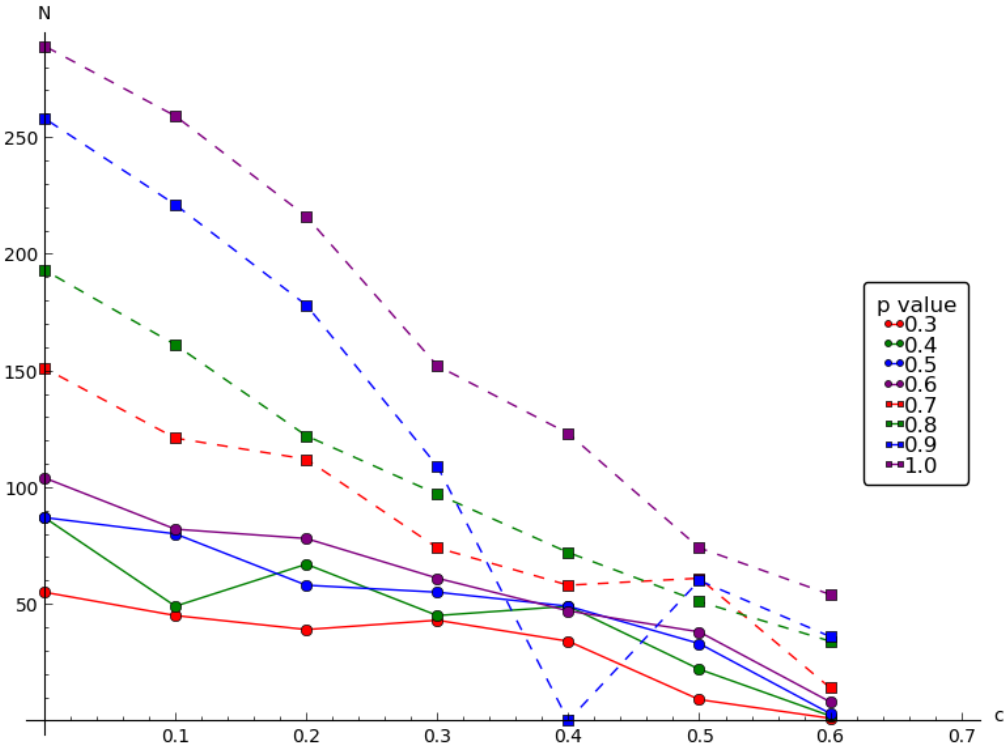
[4]Recall that the full common interest case can't even possibly have deceptive behavior, as there cannot be a sender benefit and receiver detriment at the same time.

[5]The behavior is determined in part by the representation step, though.

[6]These calculations are actually for unconscious misuse of representations, not full self-deception.

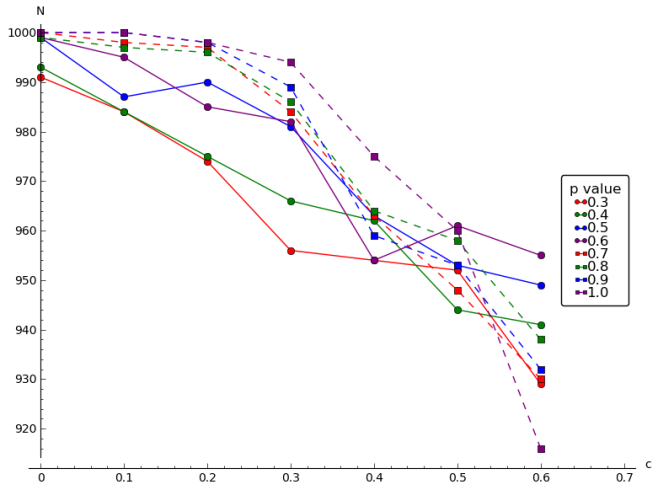


(a) Overall

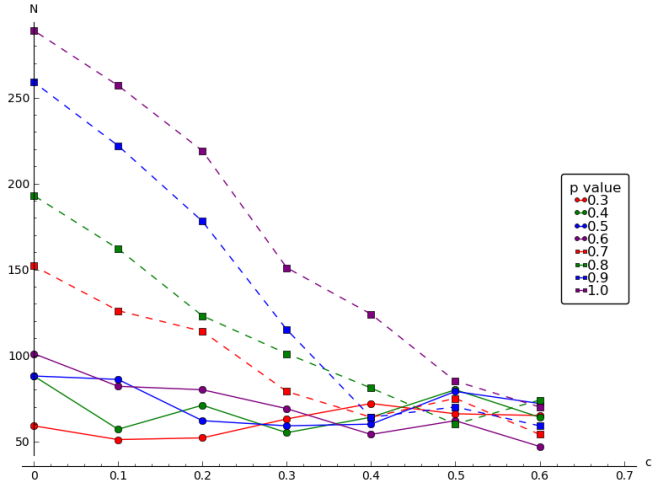


(b) Against Correct Inspection

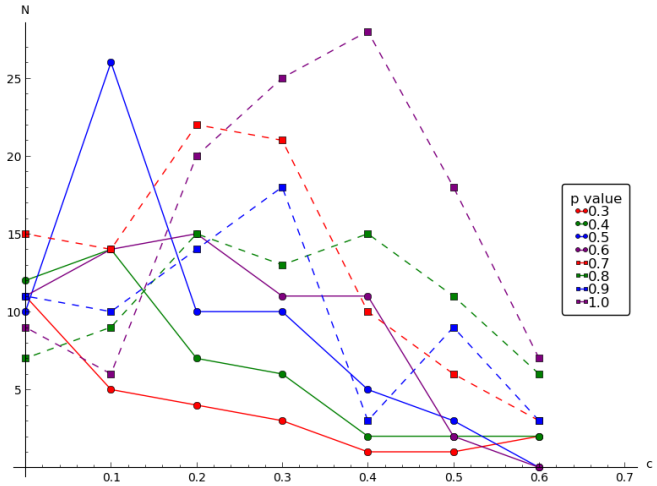
Figure 4.8: Number of Runs with Conscious Deception, $\varepsilon = 0.75$



(a) Overall



(b) Self-Deception Contributed Deception



(c) Emergent Deception

Figure 4.9: Number of Runs with Whole-Organism Deception, $\varepsilon = 0.75$

($\varepsilon = 0.9$). The level of veridical representation was substantial but noticeably less than the results for $\varepsilon = 0.75$ (Figure 4.10). Similarly, for fixed level of inspection probability, the likelihood of veridical representation increases with the cost of inspection.

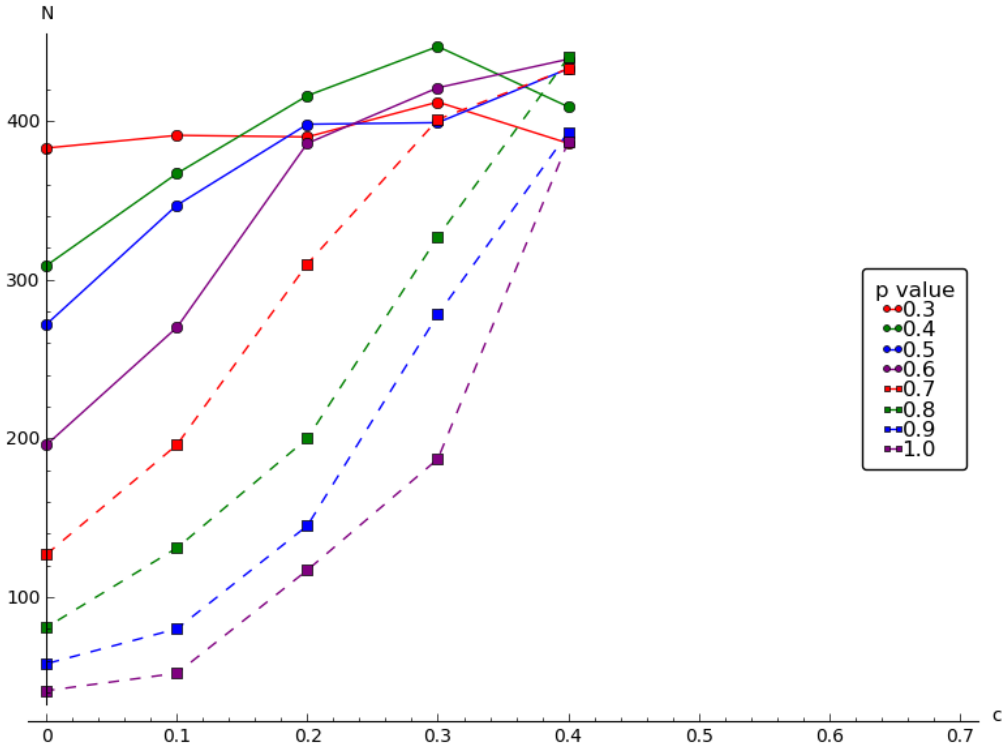
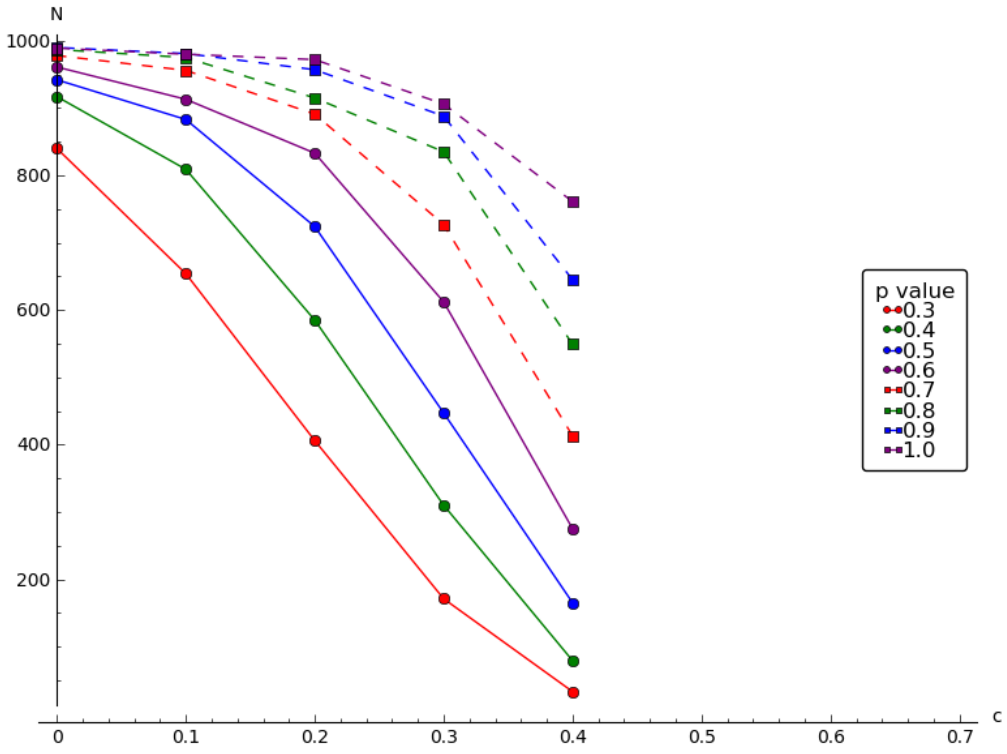


Figure 4.10: Number of Runs with Virtually Totally Veridical Representation, $\varepsilon = 0.9$

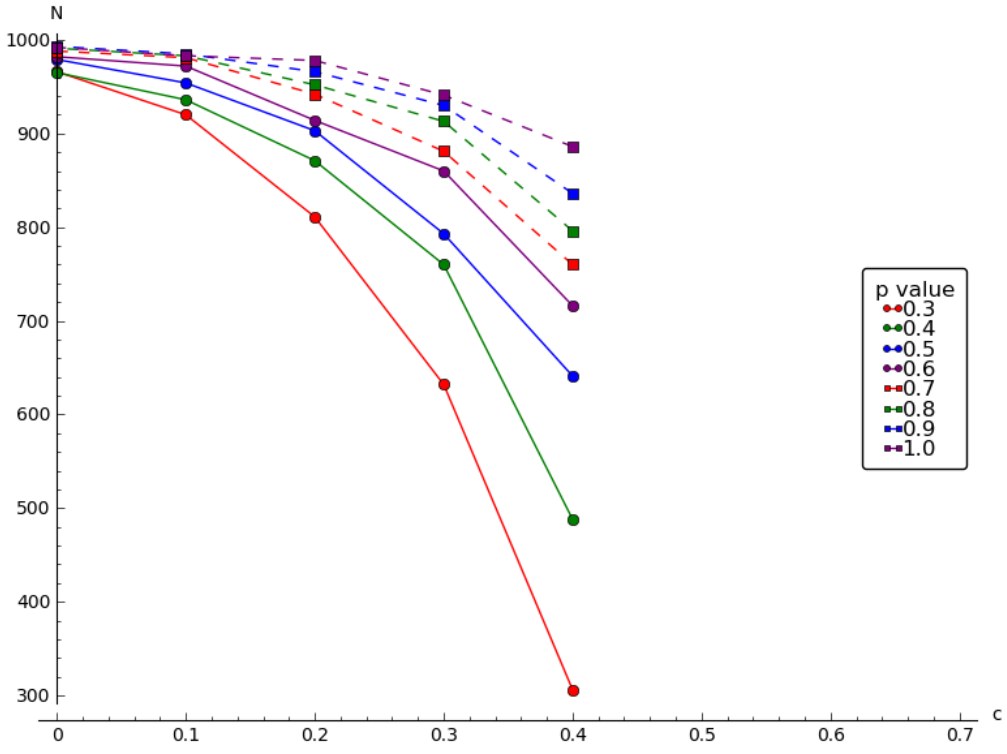
The use of the Inspect action by Receivers was again substantial, though slightly less than with $\varepsilon = 0.75$.. In nearly all cases, both with full common interest and partial common interest, the Receiver chose to Inspect at least 1% of the time (Figure 4.11). The actual percentages of inspection were also quite substantial, decreasing as the cost of inspection rose (Figure 4.12).

With less inherent incentive to represent the world veridically, the Unconscious senders in this case engaged in self-deceptive behavior much more frequently. The Unconscious was deceptive an average of 72 times with common interest and 92 times with partial common interest when $\varepsilon = 0.75$.[7] With $\varepsilon = 0.9$, those averages rose to 378 and 552 times respectively.

[7]These numbers are different from those listed previously as they are restricted to the common parameter

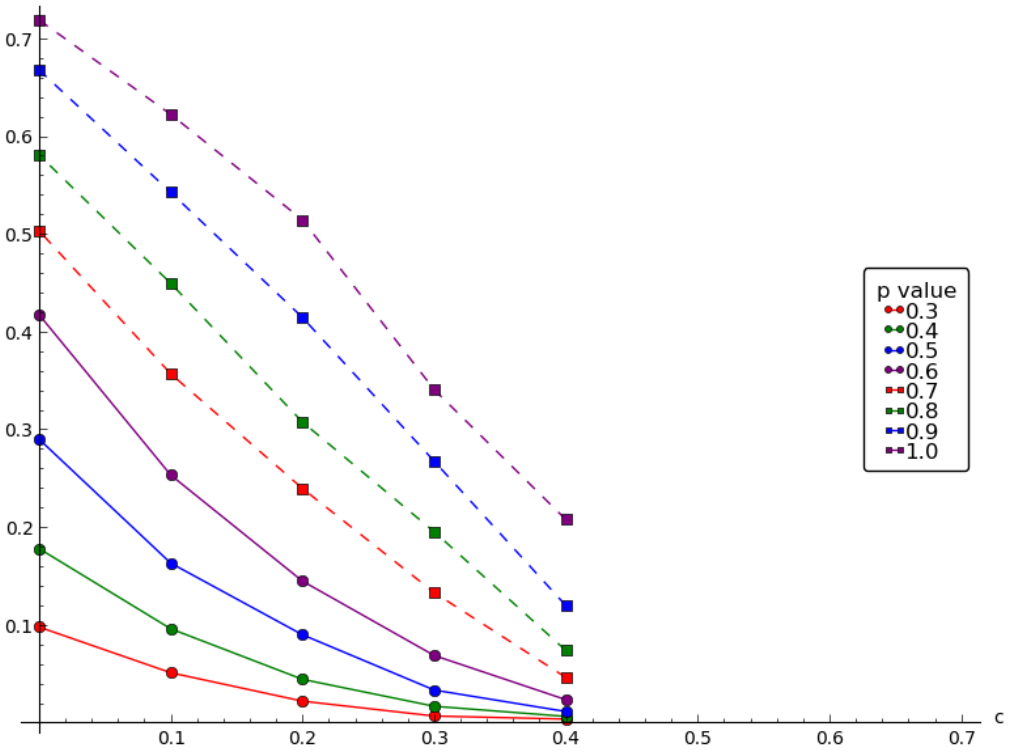


(a) Full Common Interest

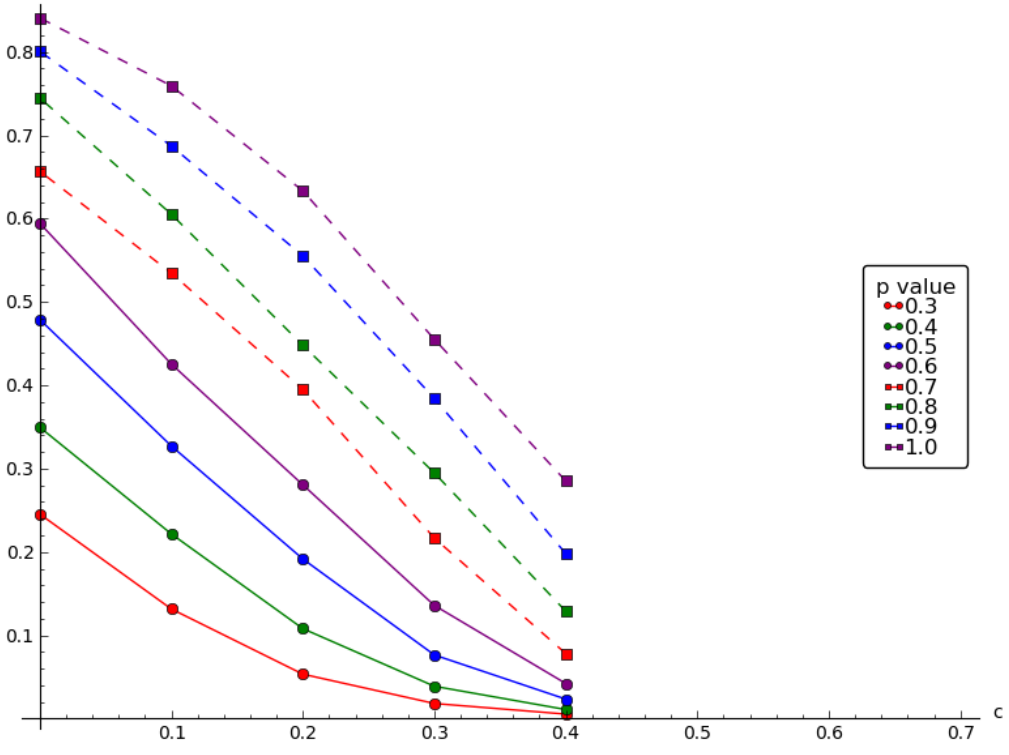


(b) Partial Common Interest

Figure 4.11: Number of Runs with Inspect Use, $\varepsilon = 0.9$



(a) Full Common Interest



(b) Partial Common Interest

Figure 4.12: Average Percentage of Inspect Use, $\varepsilon = 0.9$

Similar patterns of decrease with an increase in inspection costs were also observed (Figure 4.13).

Increased levels of self-deception in these parameter settings also corresponded with a decrease in conscious deception, when compared with the equivalent settings at $\varepsilon = 0.75$. Again, there was a downward trend for both the number of runs with conscious deception and the number of runs in which conscious deception occurred against successful inspection (Figure 4.14).

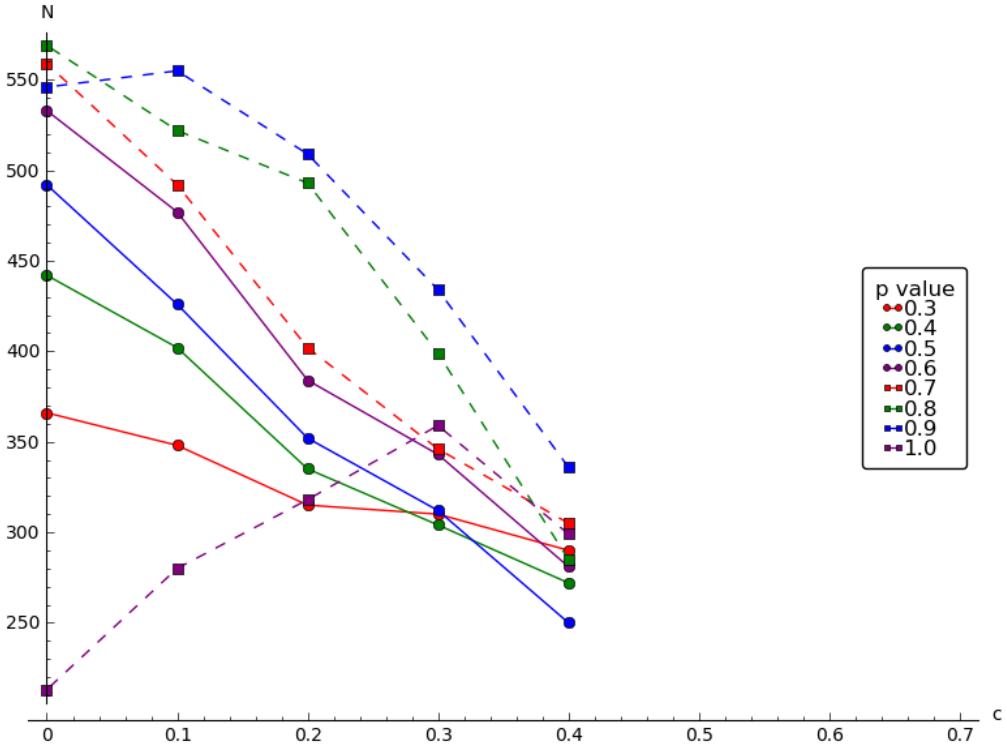
Whole-organism deception was again, as expected, more common than either conscious or self-deceptive behavior (Figure 4.21a). Many more of the runs featuring whole-organism deception did so as the result of a combination of self-deception and conscious deception (Figure 4.15b). An average of 115 runs in each parameter setting additionally featured whole-organism deception that was the result of self-deception alone. There was also an increase by a factor of about 3 or 4 in the incidence of emergent deception–deception that is neither the result of conscious deception or self-deception (or a combination of both) (Figure 4.15c).

4.4.3 Results for Epsilon at 1.0

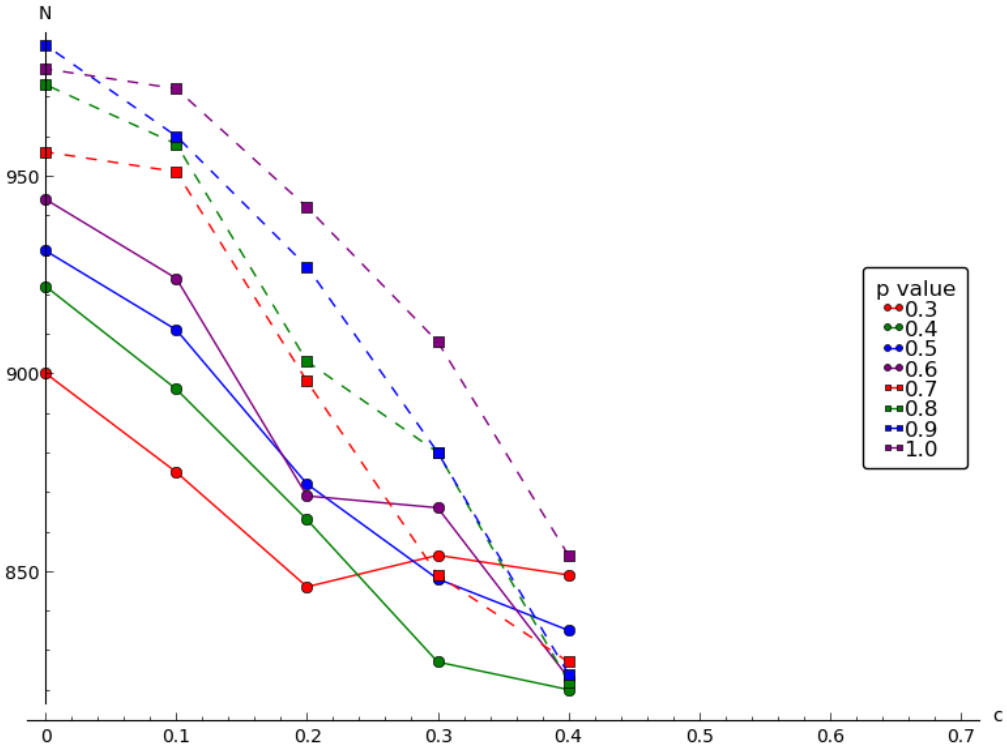
Removing the incentive for veridical representation independent of the interaction with the Receiver ($\varepsilon = 1.0$) has many of the expected effects. Veridical representations are low, approximately at chance levels (Figure 4.16).

The number of runs that featured a receiver making use of the Inspect action was surprisingly high given the lack of veridical representations (Figure 4.17). The frequency of its use, on average within each run, was comparatively low, however (Figure 4.18).

range for runs at $\varepsilon = 0.9$ so a proper comparison can be made. In particular, the settings with $c > 0.4$ were not run for $\varepsilon = 0.9$.

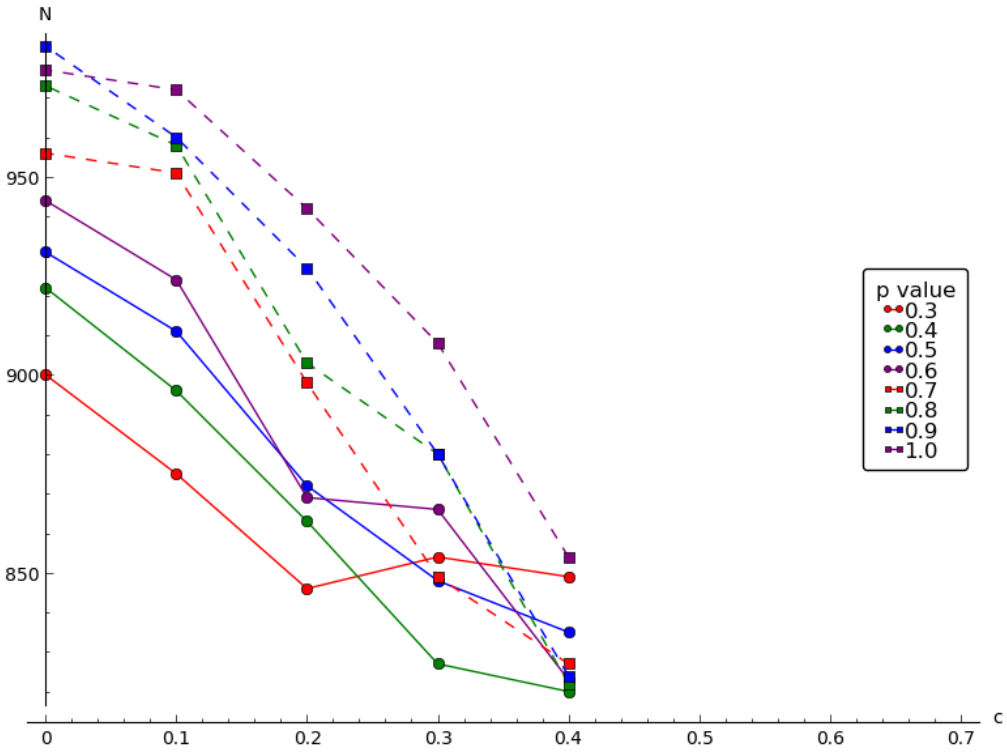


(a) Full Common Interest

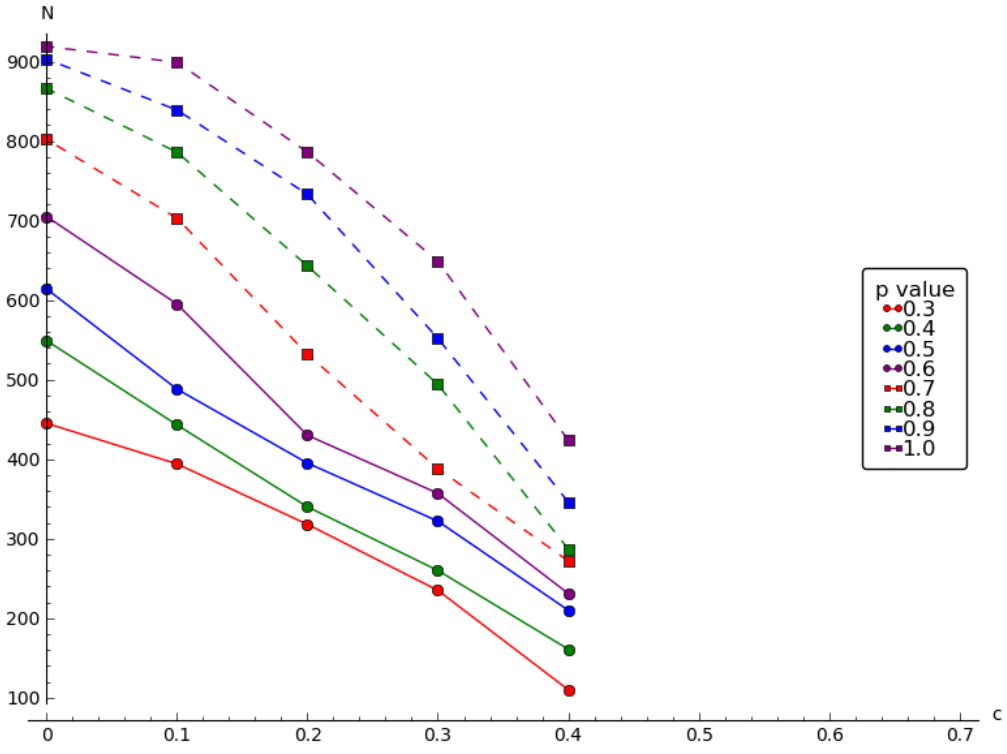


(b) Partial Common Interest

Figure 4.13: Number of Runs with Self-Deceptive Behavior, $\varepsilon = 0.9$

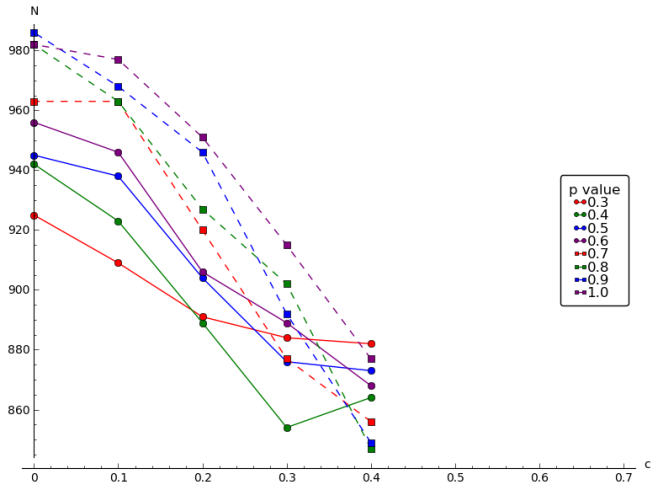


(a) Overall

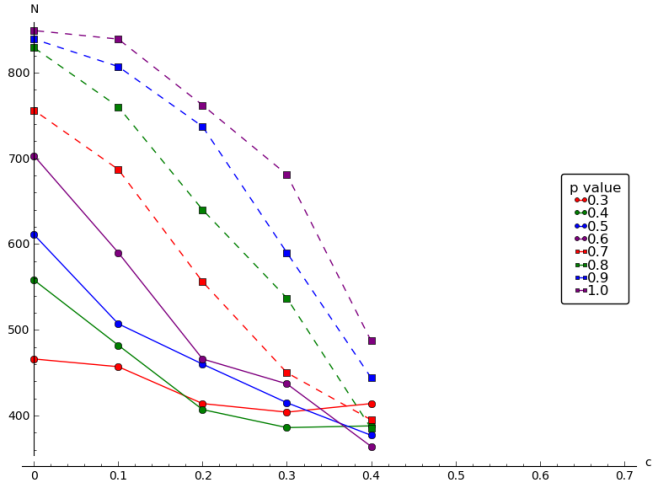


(b) Against Correct Inspection

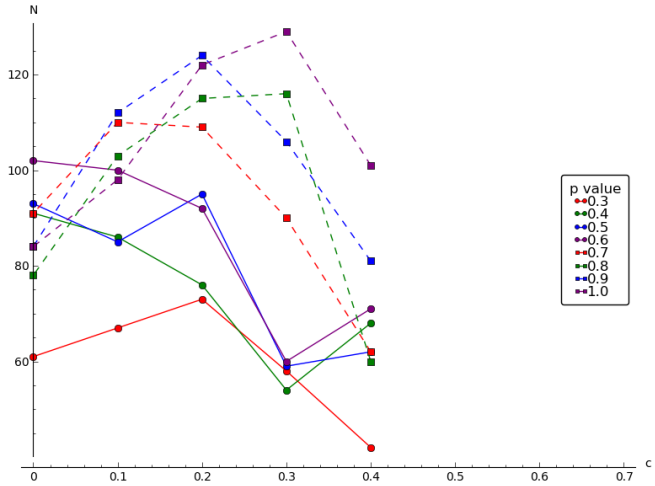
Figure 4.14: Number of Runs with Conscious Deception, $\varepsilon = 0.9$



(a) Overall



(b) Self-Deception Contributed Deception



(c) Emergent Deception

Figure 4.15: Number of Runs with Whole-Organism Deception, $\varepsilon = 0.9$

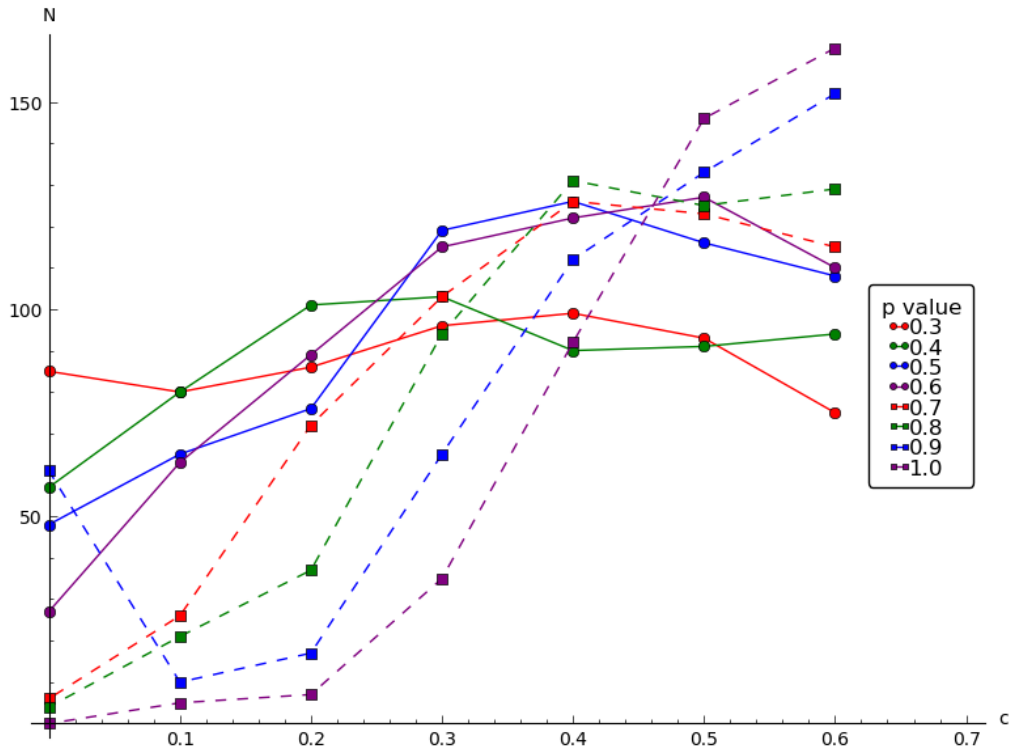
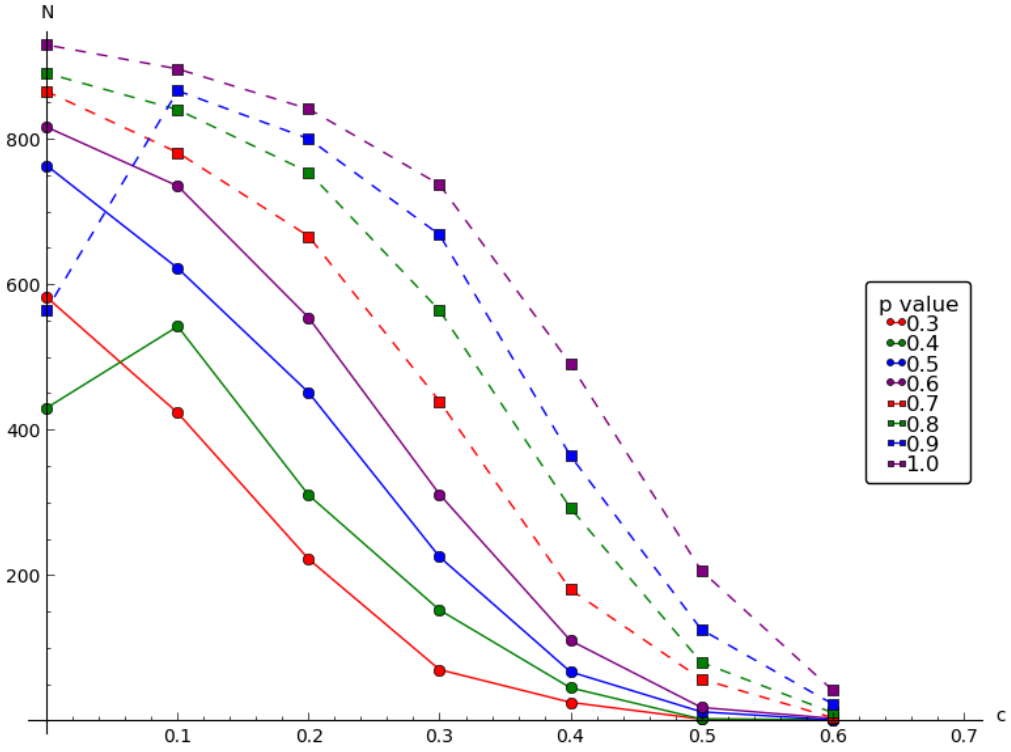


Figure 4.16: Number of Runs with Virtually Totally Veridical Representation, $\varepsilon = 1.0$

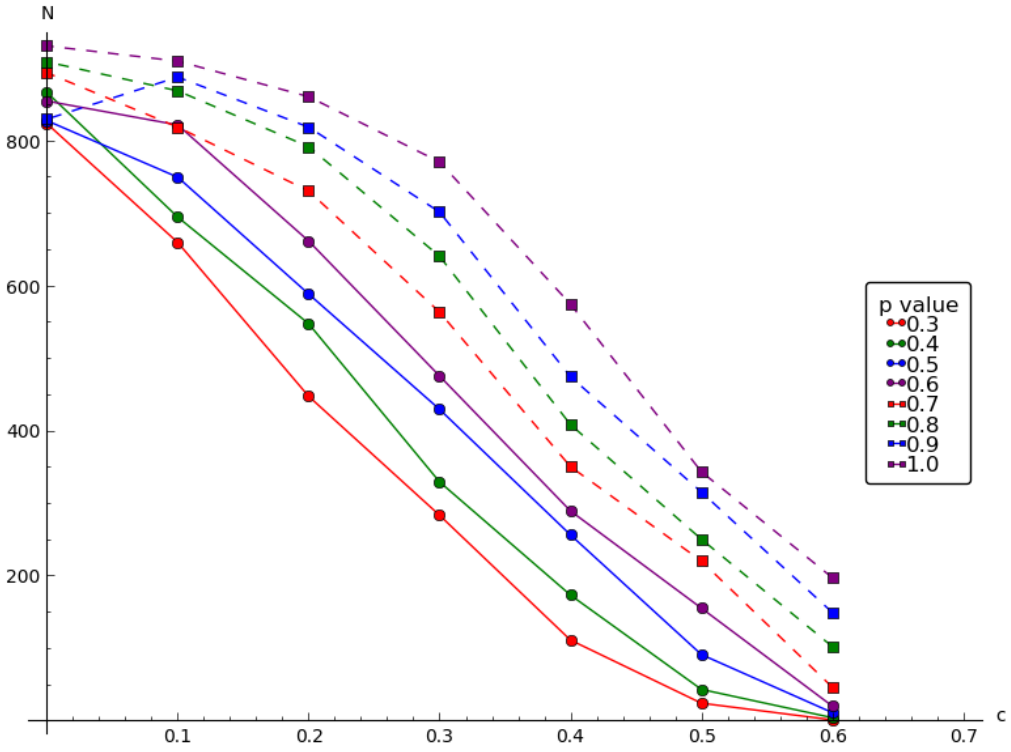
Given that veridical representation was relatively rare, self-deception levels were more substantial than they were for $\varepsilon = 0.75$, but not as large as they were for $\varepsilon = 0.9$, especially in the case of only partial common interest (Figure 4.19).

Conscious and whole-organism deception levels were high in the case of partial common interest, as they were for the other parameter settings (Figures 4.20a and 4.21a). Self-deception contributed noticeably to the levels of whole-organism deception, but not as much as it did in the case of $\varepsilon = 0.9$.[8] The levels of emergent deception were again rather low, but non-zero (Figure 4.21c).

[8]Again, self-deception here is actually just self-misuse of representations.

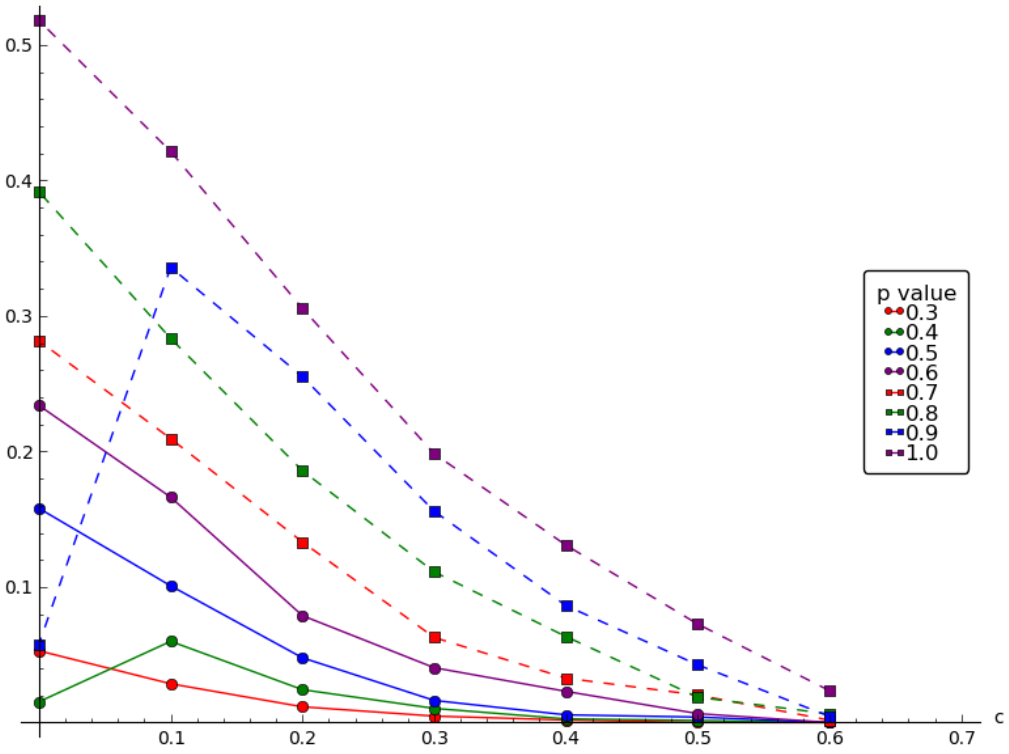


(a) Full Common Interest

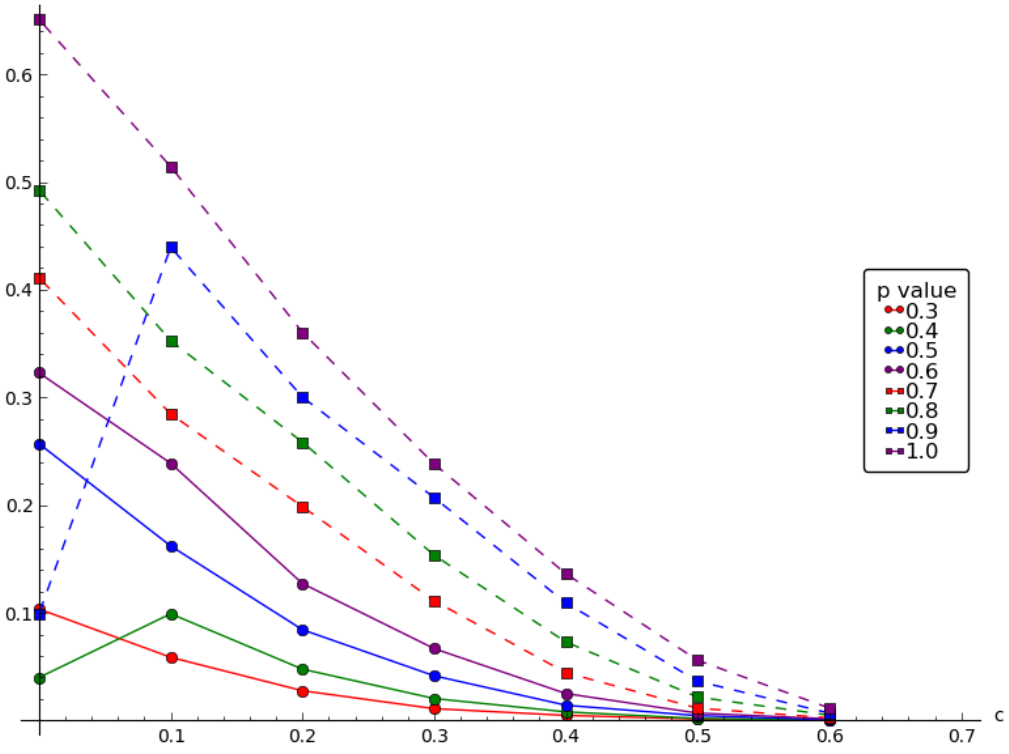


(b) Partial Common Interest

Figure 4.17: Number of Runs with Inspect Use, $\varepsilon = 1.0$

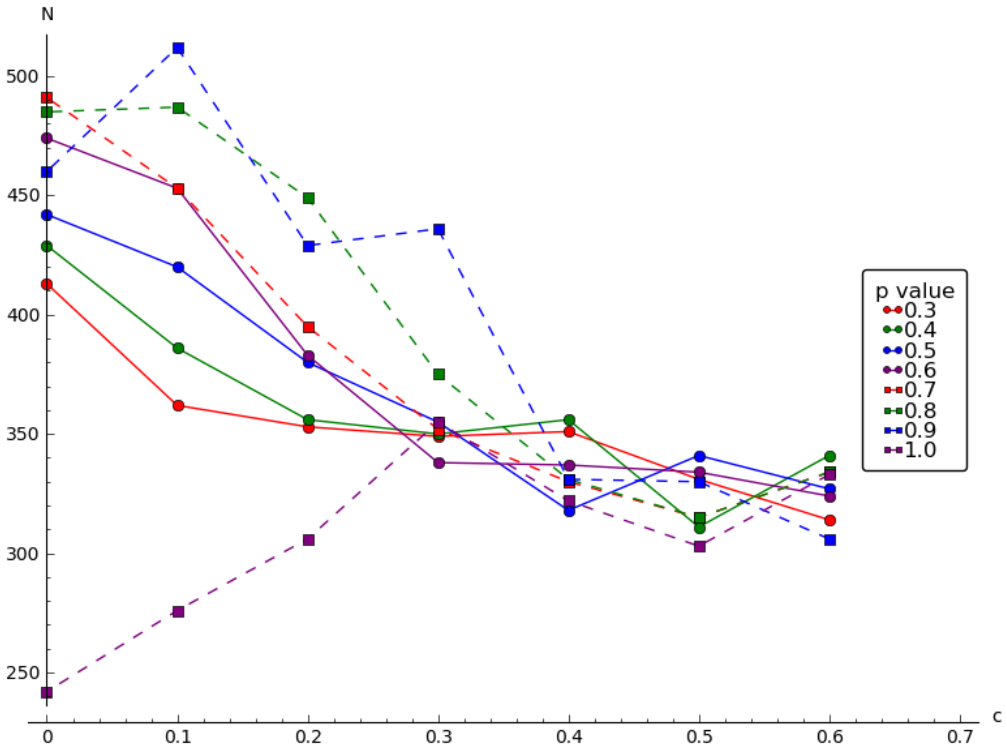


(a) Full Common Interest

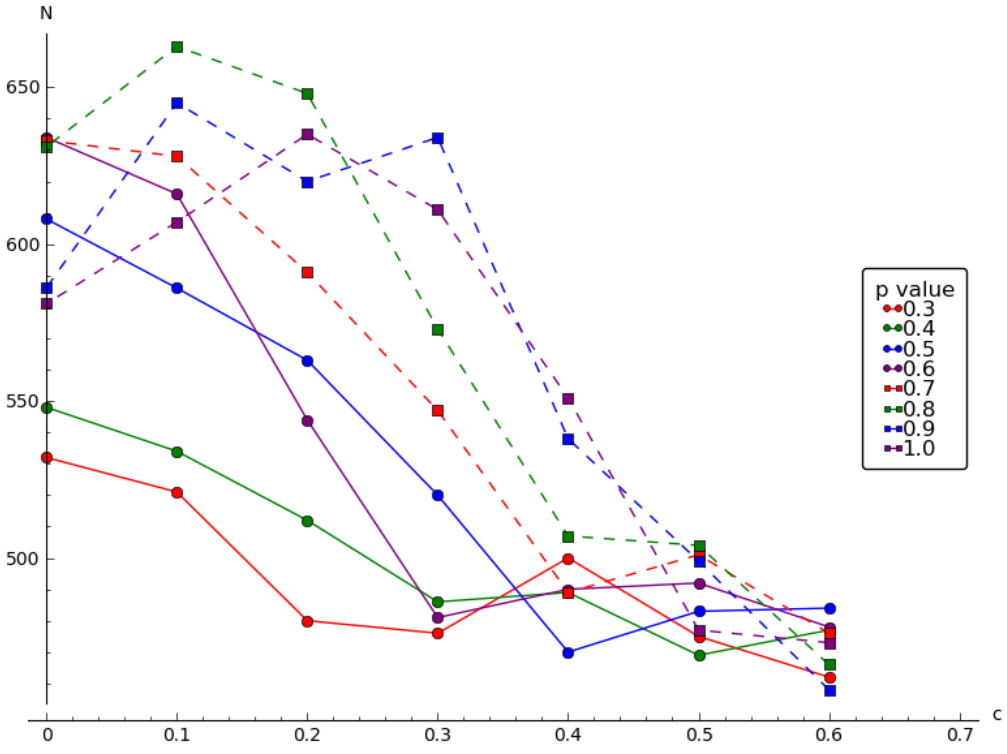


(b) Partial Common Interest

Figure 4.18: Average Percentage of Inspect Use, $\varepsilon = 1.0$

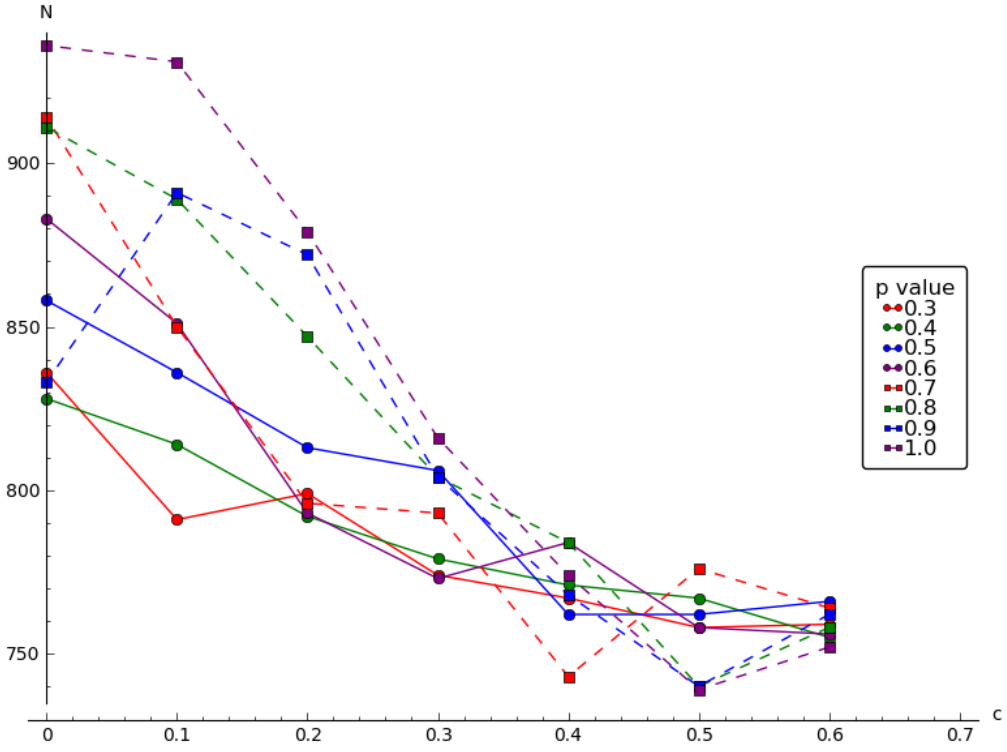


(a) Full Common Interest

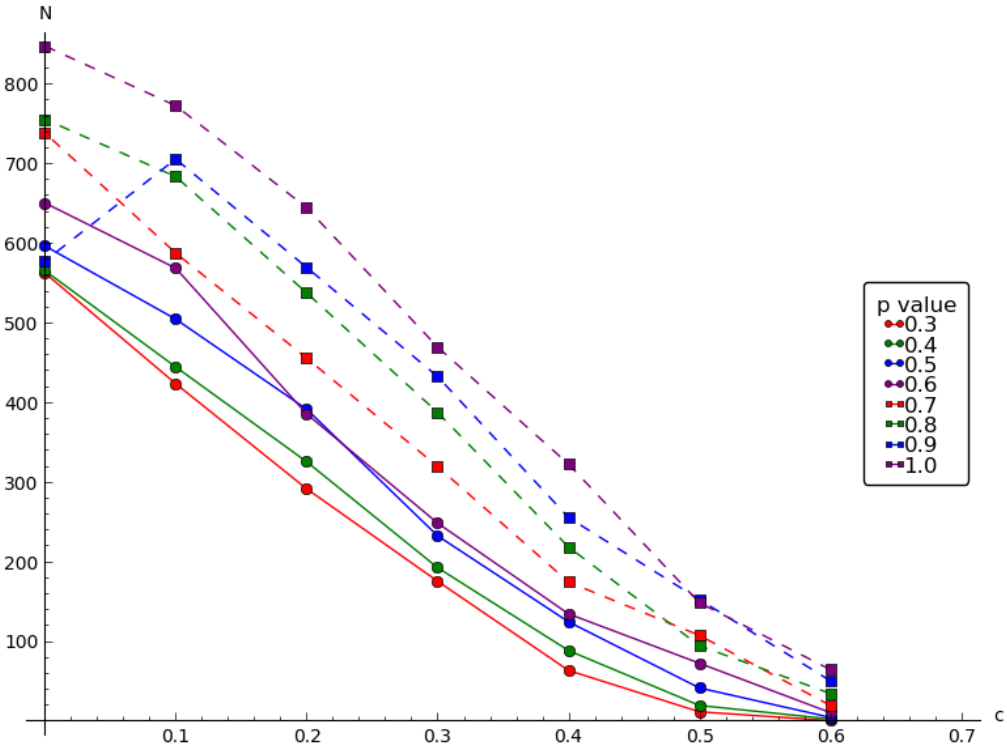


(b) Partial Common Interest

Figure 4.19: Number of Runs with Self-Deceptive Behavior, $\varepsilon = 1.0$

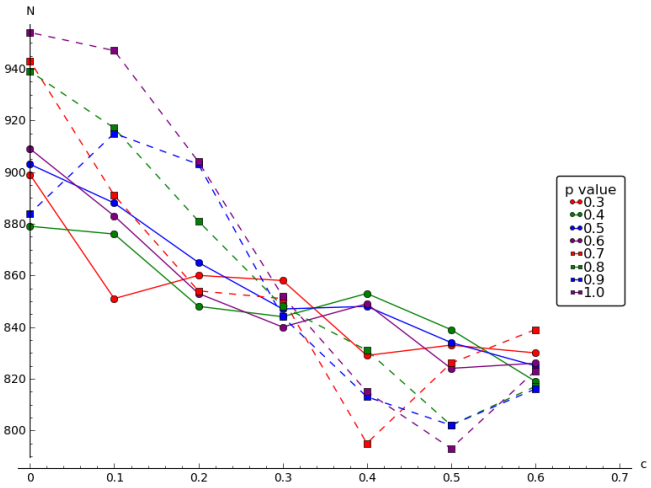


(a) Overall

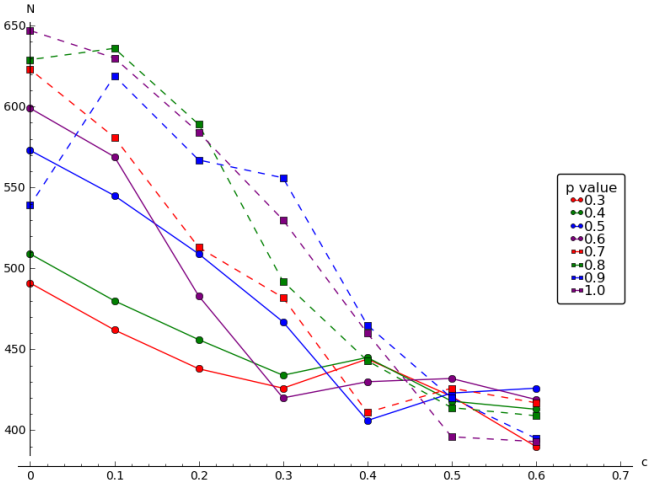


(b) Against Correct Inspection

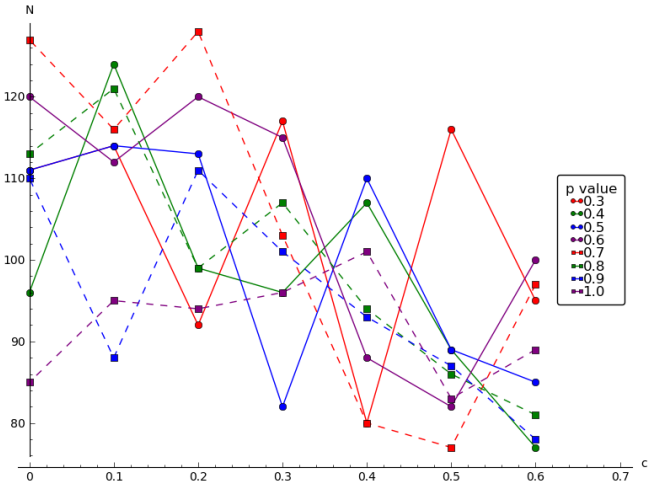
Figure 4.20: Number of Runs with Conscious Deception, $\varepsilon = 1.0$



(a) Overall



(b) Self-Deception Contributed Deception



(c) Emergent Deception

Figure 4.21: Number of Runs with Whole-Organism Deception, $\varepsilon = 1.0$

4.5 Discussion

Several questions might be raised due to the results of this model. First, there is the question of whether it is reasonable, even in humans, to think that the unconscious and conscious minds are separated enough to make the internal signalling model appropriate. Trivers thinks this is certainly possible.

> It is easy to imagine that information reaching our brain is immediately registered in consciousness and likewise that signals to initiate activity originate in the conscious mind. Of course, unconscious processes go on at the same time and unconscious processes might affect the conscious mind, but there is not a great deal of time, for example, for something like denial to operate, certainly if this requires spotting a signal and then, before it can reach consciousness, shunting it aside.... This is the conventional (pre-Freudian) view.
>
> Thirty years of accumulating evidence from neurophysiology suggests that this is an illusion... while it takes a nervous signal only about 20 ms to reach the brain, it requires a full 500 ms for a signal reaching the brain to register in consciousness! This is all the time in the world, so to speak, for emendations, changes, deletions, and enhancements to occur. Indeed neurophysiologists have shown that stimuli, at least as late as 100 ms before an occurrence reaches the consciousness, can affect the content of the experience. (Trivers, 2000, p. 119)

These timing results do seem to indicate that there is at least potentially enough time for something like an internal signalling interaction to take place. How faithfully this particular model represents the internal interactions of an agent is still a problem, but the kinds of interactions detailed in the model are at least possible.

The more important question raised by the simulation results is how plausible Trivers's story is. Might self-deception have, in fact, evolved to facilitate other-deception?

The simulation results are less than conclusive on the issue. Perhaps the most reasonable parameter settings are some of those where $\varepsilon = 0.9$. There is some incentive for the Unconscious to veridically represent the states of the world to the Conscious, but it isn't an overriding factor.

In those cases, self-deception contributing to other-deception was rather common, but conscious deception was also involved most of the time. There were some cases of self-deception alone driving deception from the whole organism, however (Figure 4.22), as well as those cases where self-misrepresentation combined with Conscious behavior resulted in whole-organism deception without either of the components being deceptive themselves.

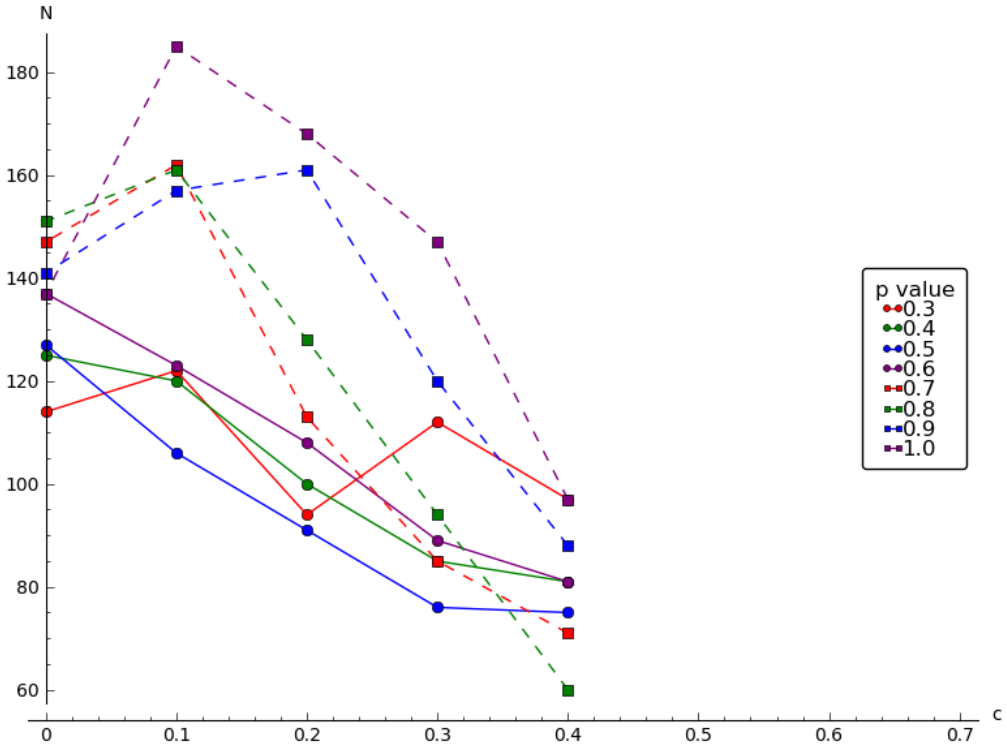


Figure 4.22: Number of Runs with Whole-organism Deception Derived Solely from Self-Deception, $\varepsilon = 0.9$

In the end, whether self-deception did in fact evolve to aid other deception seems to depend quite substantially on the factors influencing the payoffs to each agent. If making correct decisions regarding the state of the world is quite important ($\varepsilon = 0.75$, for instance), the simulations showed that there was little room for self-deception. If, on the other hand, there is incentive for veridical representation, then there is little incentive for receivers to try and discover the representation observed by the Conscious. Without such inspection, there is less reason to misrepresent the state of the world.

When the parameters do line up–enough incentive to represent veridically, but not too much; costs of inspection not too high with a reasonable probability of success–there does seem to be room for self-deception to enable other-deception.

Chapter 5

Conclusion

As the previous chapters have shown, having a good definition of deception for formal contexts can help illuminate a number of philosophical issues. It can be used to discuss claims about the possibility of universal deception, even if those claims are only inspired by (and not found in) classic texts.

It can also be used to investigate relatively new claims about the possible advantages of self-deception. The conclusions I have reached on that point are far from conclusive. Additional simulations or refinements of the model I have suggested might still bring new insights.

A potential problem is that the definition I have proposed and its applications do depend on modelling choices. More specifically, whether a situation is properly identified as deceptive or not depends on the features and quality of the model. Many situations can be modelled in more or less detail, and that detail could very possibly affect the conclusions drawn. To date, I am not aware of a good theory of modelling for signalling games. Given the proliferation of signalling models in biology, it seems that a theory of these kinds of models (concerning which might be better than others, what features are most important to capture, etc.) would be a profitable line for future research.

I don't think the lack of such a theory undermines the work I've presented in the earlier chapters. The problem I have outlined exists just as much for any other attempt to employ signalling games in the identification of deceptive behavior. The definition I have proposed remains better than the other options. Further investigating the qualities that make a model appropriate for investigating deception in a given situation would ensure that any definition of deception that employs signalling games can be unambiguously applied.

Bibliography

Bergstrom, C. T. and Lachmann, M. (1997). "Signalling among relatives. I. Is costly signalling too costly?" In: *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 352.1353, pp. 609–617.

— (1998). "Signalling among relatives. III. Talk is cheap." In: *Proceedings of the National Academy of Sciences of the United States of America* 95.9, pp. 5100–5.

Bond, C. F. and Robinson, M. (1988). "The evolution of deception." In: *Journal of Nonverbal Behavior* 12.4, pp. 295–307.

Cheney, D. L. and Seyfarth, R. M. (1992). *How Monkeys See the World: Inside the Mind of Another Species*. Chicago: University of Chicago Press.

— (2008). *Baboon Metaphysics: The Evolution of a Social Mind*. Chicago: University of Chicago Press.

Davidson, D. (1973). "Radical Interpretation." In: *Dialectica* 27.3–4, pp. 313–328.

— (1998). "Who is Fooled?" In: *Self-Deception and Paradoxes of Rationality*. Ed. by J.-P. Dupuy. Stanford: CSLI Publications, pp. 1–18.

Fingarette, H. (1969). *Self-Deception*. Berkeley, CA: University of California Press.

Godfrey-Smith, P. (2012). "Signals: Evolution, Learning, and Information, by Brian Skyrms." In: *Mind* 120.480, pp. 1288–1297.

Harms, W. (2004). "Primitive content, translation, and the emergence of meaning in animal communication." In: *Evolution of Communication Systems: A Comparative Approach*. Ed. by D. K. Oller and U. Griebel. Boston, MA: MIT Press, pp. 31–48.

Herrnstein, R. J. (1970). "On the law of effect." In: *Journal of the Experimental Analysis of Behavior* 13.2, pp. 243–266.

Huttegger, S. M. and Zollman, K. J. S. (2010). "Dynamic stability and basins of attraction in the Sir Philip Sidney game." In: *Proceedings. Biological sciences / The Royal Society* 277.1689, pp. 1915–22.

Johnstone, R. A. and Grafen, A. (1993). "Dishonesty and the handicap principle." In: *Animal Behaviour* 46.4, pp. 759–764.

Kant, I. (1999a). "Critique of Practical Reason." In: *Practical Philosophy*. Ed. by M. J. Gregor. Cambridge University Press.

— (1999b). "Groundwork of the Metaphysics of Morals." In: *Practical Philosophy*. Ed. by M. J. Gregor. Cambridge University Press.

— (1999c). "On a supposed right to lie from philanthropy." In: *Practical Philosophy*. Ed. by M. J. Gregor. Cambridge University Press.

Kullback, S. and Leibler, R. (1951). "On information and sufficiency." In: *The Annals of Mathematical Statistics* 22.1, pp. 79–86.

Lachmann, M. and Bergstrom, C. T. (1998). "Signalling among relatives II. Beyond the Tower of Babel." In: *Theoretical Population Biology* 54.2, pp. 146–160.

— (2004). "The Disadvantage of Combinatorial Communication." In: *Proceedings of the Royal Society of London. Series B, Biological Sciences* 271, pp. 2337–2343.

Lewis, D. (1969). *Convention: A Philosophical Study*. Oxford: Blackwell.

Maynard Smith, J. (1991). "Honest signalling: the Philip Sidney game." In: *Animal Behaviour* 42, pp. 1034–1035.

Maynard Smith, J. and Harper, D. (2003). *Animal Signals*. Oxford: Oxford University Press.

Millikan, R. G. (2004). "On Reading Signs: Some Differences Between Us and the Others." In: *Evolution of Communication Systems: A Comparative Approach*. Ed. by D. K. Oller and U. Griebel. Boston, MA: MIT Press, pp. 15–29.

Mitchell, R. W. (1986). "A Framework for Discussing Deception." In: *Deception. Perspectives on Human and Nonhuman Deceit*. Ed. by R. W. Mitchell and N. S. Thompson. Albany: State University of New York Press, pp. 3–40.

Nöldeke, G. and Samuelson, L. (2003). "Strategic Choice Handicaps When Females Seek High Male Net Viability." In: *Journal of Theoretical Biology* 221.1, pp. 53–59.

Qunie, W. v. O. (1960). *Word & Object*. Boston, MA: MIT Press.

Reaka, M. (1979). "Patterns of molting frequencies in coral-dwelling stomatopod Crustacea." In: *The Biological Bulletin* 156.3, pp. 328–342.

Roth, A. E. and Erev, I. (1995). "Learning in extensive-form games: Experimental data and simple dynamic models in the intermediate term." In: *Games and Economic Behavior* 8.1, pp. 164–212.

Searcy, W. A. and Nowicki, S. (2005). *The Evolution of Animal Communication. Reliability and Deception in Signaling Systems*. Princeton: Princeton University Press.

Semple, S. and McComb, K. (1996). "Behavioural deception." In: *Trends in Ecology and Evolution* 11.10, pp. 434–437.

Skyrms, B. (2010). *Signals. Evolution, Learning, & Information*. Oxford University Press.

Steger, R. and Caldwell, R. (1983). "Intraspecific deception by bluffing: a defense strategy of newly molted stomatopods (Arthropoda: Crustacea)." In: *Science* 221.4610, pp. 558–560.

Sutter, M. (2009). "Deception Through Telling the Truth?! Experimental Evidence From Individuals and Teams." In: *The Economic Journal* 119.534, pp. 47–60.

Számadó, S. (2000). "Cheating as a mixed strategy in a simple model of aggressive communication." In: *Animal Behaviour* 59.1, pp. 221–230.

Trivers, R. (2000). "The elements of a scientific theory of self-deception." In: *Annals of the New York Academy of Sciences* 907, pp. 114–31.

Trivers, R. (2011). *The Folly of Fools*. Basic Books.

von Hippel, W. and Trivers, R. (2011). "The evolution and psychology of self-deception." In: *The Behavioral and brain sciences* 34.1, pp. 1–16, 1–16.

Weibull, J. W. (1995). *Evolutionary Game Theory*. MIT Press.

Zahavi, A. (1975). "Mate selection: a selection for a handicap." In: *Journal of theoretical Biology* 53.1, pp. 205–214.

— (1977). "The cost of honesty (further remarks on the handicap principle)." In: *Journal of Theoretical Biology* 67.3, p. 603.