Strategies of persuasion, manipulation and propaganda: psychological and social aspects

Michael Franke & Robert van Rooij

Abstract

How can one influence the behavior of others? What is a good persuasion strategy? It is obviously of great importance to determine *what* information best to provide and also *how* to convey it. To delineate how and when manipulation of others can be successful, the first part of this paper reviews basic findings of decision and game theory on models of strategic communication. But there is also a social aspect to manipulation, concerned with determining *who we should address* so as best to promote our opinion in a larger group or society as a whole. The second half of this paper therefore looks at a novel extension of DeGroot's (1974)'s classical model of opinion dynamics that allows agents to strategically influence some agents more than others. This side-by-side investigation of psychological and social aspects enables us to reflect on the general question what a good manipulation strategy is. We submit that successful manipulation requires exploiting critical weaknesses, such as limited capability of strategic reasoning, limited awareness, susceptibility to cognitive biases or to potentially indirect social pressure.

You might be an artist, politician, banker, merchant, terrorist, or, what is likely given that you are obviously reading this, a scientist. Whatever your profession or call of heart, your career depends, whether you like it or not, in substantial part on your success at influencing the behavior and opinions of others in ways favorable to you (but not necessarily favorable to them). Those who put aside ethical considerations and aspire to be successful manipulators face two major challenges. The first challenge is the most fundamental and we shall call it *pragmatic* or *one-to-one*. It arises during the most elementary form of manipulative effort whenever a single manipulator faces a single decision maker whose opinion or behavior the former seeks to influence. The one-to-one challenge is mostly, but not exclusively, about *rhetoric*, i.e., the proper use of logical arguments and other, less normatively compelling, but perhaps even more efficiently persuasive communication strategies. But if manipulation is to be taken further, also a second challenge arises and that is *social* or *many-to-many*. Supposing that we know *how* to exert efficient influence, it is another issue *who* to exert influence on in a group of decision makers, so as to efficiently propagate an opinion in a society.

This paper deals with efficient strategies for manipulation at both levels. This is not only relevant for aspiring master manipulators, but also for those who would like to brace themselves for a life in a manipulative environment. Our main conclusions are that successful manipulation requires the exploitation of weaknesses of those to be

manipulated. So in order to avoid being manipulated, it is important to be aware of the possibility of malign manipulation and one's own weaknesses.

The paper is divided into two parts. The first is addressed in Section 1 and deals with the pragmatic perspective. It first shows that standard models from decision and game theory predict that usually an ideally rational decision maker would see through any manipulative effort. But if this is so, there would not be much successful manipulation, and also not many malign persuasive attempts from other ideally rational agents. Since this verdict flies in the face of empirical evidence, we feel forced to extend our investigation to more psychologically adequate models of boundedly rational agency. Towards this end, we review models of (i) unawareness of the game/context model, (ii) depth-limited step-by-step reasoning, and (iii) descriptive decision theory. We suggest that it is cognitive shortcomings of this sort that manipulators have to exploit in order to be successful.

Whereas Section 1 has an overview character in that it summarizes key notions and insights from the relevant literature, Section 2 seeks to explore new territory. It investigates a model of social *opinion dynamics*, i.e., a model of how opinions spread and develop in a population of agents, which also allows agents to choose whom to influence and whom to neglect. Since the complexity of this social dimension of manipulation is immense, the need for simple yet efficient heuristics arises. We try to delineate in general terms what a good heuristic strategy is for social manipulation of opinions and demonstrate with a case study simulating the behavior of four concrete heuristics in different kinds of social interaction structures that (i) strategies that aim at easily influenceable targets are efficient on a short time scale, while strategies that aim at influential targets are efficient on a longer time scale, and that (ii) it helps to play a coalition strategy together with other likeminded manipulators, in particular so as not to get into one another's way. This corroborates the general conclusion that effective social propaganda, like one-to-one strategic manipulation, requires making strategic use of particularly weak spots in the flow patterns of information within a society.

Another final contribution of this paper is in what it is *not* about. To the best of our knowledge, there is little systematic work in the tradition of logic and game theory that addresses both the psychological and the social dimension of strategic manipulation at once. We therefore conclude the paper with a brief outlook at the many vexing open issues that arise when this integrative perspective is taken seriously.

A Note on Terminology. When we speak of a *strategy* here, what we have in mind is mostly a very loose and general notion, much like the use of the word "strategy" in non-technical English, when employed by speakers merrily uninterested in any geeky meaning contrast between "strategy" and "tactic". When we talk about a 'good' strategy, we mean a communication strategy that influences other agents to act, or have an opinion, in accordance with the manipulator's preferences. This notion of communication strategy is different from the one used in other contributions to this volume.

Within game theory, the standard notion of a strategy is that of a *full contingency plan* that specifies at the beginning of a game which action an agent chooses whenever she might be called to act. When we discuss strategies of games in Section 1 as a formal specification of an agent's behavior, we do too use the term in this specific

technical sense. In general, however, we talk about strategic manipulation from a more God's-eye point of view, referring to a good strategy as what is a good general principle which, if realized in a concrete situation, would give rise to a "strategy" in the formal, game theoretic sense of the term.

1 Pragmatic aspects of persuasion and manipulation

The pragmatic dimension of persuasion and manipulation chiefly concerns the use of language. Persuasive communication of this kind is studied in rhetoric, argumentation theory, politics, law, and marketing. But more recently also pragmatics, the linguistic theory of language use, has turned its eye towards persuasive communication, especially in the form of *game theoretic pragmatics*. This is a very welcome development, for two main reasons. Firstly, the aforementioned can learn from pragmatics: a widely used misleading device in advertisements—a paradigmatic example of persuasion—is *false implication* (e.g. Kahane and Cavender, 1980). A certain quality is claimed for the product without explicitly asserting its uniqueness, with the intention to make you assume that only that product has the relevant quality. Persuasion by false implication is reminiscent of *conversational implicature*, a central notion studied in linguistic pragmatics (e.g. Levinson, 1983). Secondly, the study of persuasive communication *should* really be a natural part of linguistic pragmatics. The only reason why persuasion has been neglected for long is due to the fact that the prevalent theory of language use in linguistics is based on the Gricean assumption of *cooperativity* (Grice, 1975). Though game theory can formalize Gricean pragmatics, its analysis of strategic persuasive communication is suitable for non-cooperative situations as well. Indeed, game theory is the natural framework for studying strategic manipulative communication.

To show this, the following Sections 1.1 and 1.2 introduce the main setup of decision and game-theoretic models of one-to-one communication. Unfortunately, as we will see presently, standard game theory counterintuitively predicts that successful manipulation is rare if not impossible. This is because, ideally rational agents would basically see through attempts of manipulation. Hence ideally rational manipulators would not even try to exert malign influence. In reaction to this counterintuitive predicament, Section 1.3 looks at a number of models in which some seemingly unrealistic assumptions of idealized rational agency are levelled. In particular, we briefly cover models of (i) language use among agents who are possibly unaware of relevant details of the decision-making context, (ii) language use among agents who are limited in their depth of strategic thinking, and (iii) the impact that certain surprising features and biases of our cognitive makeup, such as *framing effects* (Kahnemann and Tversky, 1973), have on decision making.

1.1 Decisions and information flow

On first thought it may seem that it is always helpful to provide truthful information and mischievous to lie. But this first impression is easily seen to be wrong. For one thing, it can sometimes be helpful to lie. For another, providing truthful but incomplete information can sometimes be harmful.

Here is a concrete example that shows this. Suppose that our decision maker is confronted with the decision problem whether to choose action a_1 or a_2, while uncertain which of the states $t_1, \ldots, t_6$ is actual:

$U(a_i, t_j)$	t_1	t_2	t_3	t_4	t_5	t_6
a_1	-1	1	3	7	-1	1
a_2	2	2	2	2	2	2

By definition, rational decision makers choose their actions so as to maximize their expected utility. So, if a rational agent considers each state equally probable, it is predicted that he will choose a_2 because that has a higher expected utility than a_1: a_2 gives a sure outcome of 2, but a_1 only gives an expected utility of $5/3 = 1/6 \times \sum_i u(a_1, t_i)$. If t_1 is the actual state, the decision maker has made the right decision. This is not the case, however, if, for instance, t_3 were the actual state. It it is now helpful for the decision maker to receive the false information that t_4 is the actual state: falsely believing that t_4 is actual, the decision maker would choose the action which is in fact best in the actual state t_3. And of course, we all make occasional use of *white lies*: communicating something that is false in the interest of tact or politeness.

Another possibility is providing truthful but misleading information. Suppose that the agent receives the information that states t_5 and t_6 are not the case. After updating her information state (i.e., probability function) by standard conditionalization, rationality now dictates our decision maker to choose a_1 because that now has the highest expected utility: $5/2$ versus 2. Although a_1 was perhaps the most rational action to choose given the decision maker's uncertainty, he still made the *wrong decision* if it turns out that t_1 is the actual state. One can conclude that receiving truthful information is not always helpful, and can sometimes even hurt.

Communication helps to disseminate information. In many cases, receiving truthful information is helpful: it allows one to make a better informed decision. But we have just seen that getting truthful information can be harmfull as well, at least when it is partial information. As a consequence, there is room for maling manipulation even with the strategic dissemination of truthful information, unless the decision maker would realize the potentially intended deception. Suppose, for instance, that the manipulator prefers the decision maker to perform a_1 instead of a_2, independently of which state actually holds. If the decision maker and the manipulator are both ideally rational, the informer will realize that it doesn't make sense to provide, say, information $\{t_1, t_2, t_3, t_4\}$ with misleading intention, because the decision maker won't fall for this and will consider information to be *incredible*. A new question comes up: how much can an agent credibly communicate in a situation like that above? This type of question is studied by economists making use of signaling games.

1.2 Signaling games and credible communication

Signaling games are the perhaps simplest non-trivial game-theoretic models of language use. They were invented by David Lewis to study the emergence of conventional semantic meaning (Lewis, 1969). For reasons of exposition, we first look at Lewisean

signaling games where messages do not have a previously given conventional meaning, but then zoom in on the case where a commonly known conventional language exists.

A signaling game proceeds as follows. A sender S observes the actual state of the world $t \in T$ and chooses a message m from a set of alternatives M. In turn, R observes the sent message and chooses an action a from a given set A. The payoffs for both S and R depend in general on the state t, the sent message m and the action a chosen by the receiver. Formally, a *signaling game* is a tuple $\langle \{S, R\}, T, \Pr, M, A, \mathrm{U}_S, \mathrm{U}_R \rangle$ where $\Pr \in \Delta(T)$ is a probability distribution over T capturing the receiver's *prior beliefs* about which state is actual, and $\mathrm{U}_{S,R} : M \times A \times T \to \mathbb{R}$ are utility functions for both sender and receiver. We speak of a *cheap-talk game*, if message use does not influence utilities.[1]

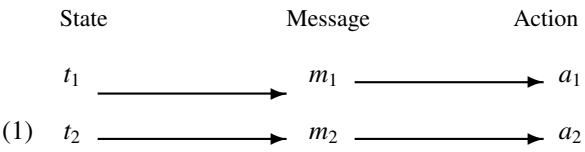
It is clear to see that a signaling game embeds a classical decision problem, such as discussed in the previous section. The receiver is the decision maker and the sender is the manipulator. It is these structures that help us to study manipulation strategies and assess their success probabilities.

To specify player behavior, we define the notion of a *strategy*. (This is now a technical use of the term, in line with the remarks above.) A *sender strategy*, $\sigma \in M^T$ is modelled as a function from states to messages. Likewise, a *receiver strategy* $\rho \in A^M$ is a function from messages to actions. The strategy pair $\langle \sigma^*, \rho^* \rangle$ is an equilibrium if neither player can do any better by unilateral deviation. More technically, $\langle \sigma^*, \rho^* \rangle$ is a *Nash equilibrium* iff for all $t \in T$:

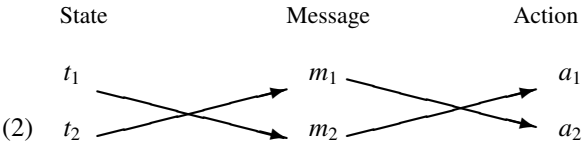
(i) $\mathrm{U}_S(t, \sigma^*(t), \rho^*(\sigma^*(t))) \geq \mathrm{U}_S(t, \sigma(t), \rho^*(\sigma(t)))$ for all $\sigma \in M^T$, and

(ii) $\mathrm{U}_R(t, \sigma^*(t), \rho^*(\sigma^*(t))) \geq \mathrm{U}_R(t, \sigma^*(t), \rho(\sigma^*(t)))$ for all $\rho \in A^M$.

A signaling game typically has many equilibria. Suppose we limit ourselves to a cooperative signaling game with only two states $T = \{t_1, t_2\}$ that are equally probable $\Pr(t_1) = \Pr(t_2)$, two messages $M = \{m_1, m_2\}$, and two actions $A = \{a_1, a_2\}$, and where $\mathrm{U}(t_i, m_j, a_k) = 1$ if, $i = k$, and 0 otherwise, for both sender and receiver. In that case the following combination of strategies is obviously a Nash equilibrium:[2]



The following combination of strategies is an equally good equilibrium:

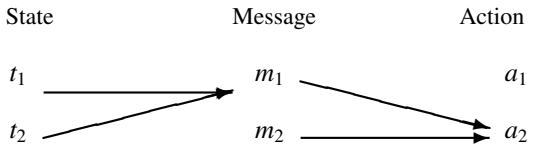


[1] For simplicity we assume that T, M and A are finite non-empty sets, and that $\Pr(t) > 0$ for all $t \in T$.

[2] Arrows from states to messages depict sender strategies; arrows from messages to actions depict receiver strategies.

In both situations, the equilibria make real communication possible. Unfortunately, there are also Nash equilibria where nothing is communicated about the actual state of affairs. In case the sender's prior probability of t_2 exceeds that of t_1, for instance, the following combination is also a Nash equilibrium:

(3)



Until now we assumed that messages don't have an *a priori* given meaning. What happens if we give up this assumption, and assume that a conventional language is already in place that can be used or abused by speakers to influence their hearers for better or worse? Formally, we model this by a semantic denotation function $[\![\cdot]\!] : M \rightarrow \mathcal{P}(T)$ such that $t \in [\![m]\!]$ iff m is true in t.[3]

Assuming that messages have a conventional meaning can help filter out unreasonable equilibria. In seminal early work, Farrell (1993) (the paper goes back to at least 1984) proposed to refine the equilibrium set for cheap-talk signaling games by a notion of *message credibility*, requiring that R believe what S says if it is in S's interest to speak the truth (c.f. Farrell and Rabin, 1996). Farrell's solution is rather technical and can be criticized for being unrealistic, but his general idea has been picked up and refined in many subsequent contributions, as we will also see below (c.f. Myerson, 1989; Rabin, 1990; Matthews et al., 1991; Zapater, 1997; Stalnaker, 2006; Franke, 2010). Essentially, Farrell assumed that the set of available messages is infinite and expressively rich: for any given reference equilibrium and every subset $X \subseteq T$ of states, there is always a message m_X with $[\![m_X]\!] = X$ that is not used in that equilibrium.[4] Such an unused message m is called a *credible neologism* if, roughly speaking, it can overturn a given reference equilibrium. Concretely, take an equilibrium $\langle \sigma^*, \rho^* \rangle$, and let $U_S^*(t)$ be the equilibrium payoff of type t for the sender. The types in $[\![m]\!]$ can send a *credible neologism* iff $[\![m]\!] = \{t \in T : U_S(t, BR([\![m]\!])) > U_S^*(t)\}$, where $BR([\![m]\!])$ is R's (assumed unique, for simplicity) optimal response to the prior distribution conditioned on $[\![m]\!]$. If R interprets a credible neologism literally, then some types would send the neologism and destroy the candidate equilibrium. A *neologism proof equilibrium* is an equilibrium for which no subset of T can send a credible neologism. For example, the previous two fully revealing equilibria in (1) and (2) are neologism proof, but the pooling equilibrium in (3) is not: there is a message m^* with $[\![m^*]\!] = \{t_2\}$ which only t_2 would prefer to send over the given pooling equilibrium.

Farrell defined his notion of credibility in terms of a given reference equilibrium. Yet for accounts of online pragmatic reasoning about language use, it is not always clear where such an equilibrium should come from. In that case another reference point for pragmatic reasoning is ready-at-hand, namely a situation *without* communication

[3] We assume for simplicity that for each state t there is at least one message m which is true in that state; and that no message is contradictory, i.e., there is no m for which $[\![m]\!] = \emptyset$.

[4] This *rich language assumption* might be motivated by evolutionary considerations, but is unsuitable for applications to online pragmatic reasoning about natural language, which, arguably, is not at the same time cheap and fully expressive: some things are more cumbersome to express than others (c.f. Franke, 2010).

entirely. So another way of thinking about $U^*_S(t)$ is just as the utility of S in t if R plays the action with the highest expected utility of R's decision problem. In this spirit, van Rooy (2003) determines the *relevance of information* against the background of the decision maker's decision problem. Roughly speaking, the idea is that message m is relevant w.r.t. a decision problem if the hearer will change his action upon hearing it.[5] A message is considered credible in case it is relevant, and cannot be used misleadingly. As an example, let's look at the following cooperative situation:

(4)

$U(t_i, a_j)$	a_1	a_2
t_1	1,1	0,0
t_2	0,0	1,1

If this was just a decision problem without possibility of communication and furthermore $Pr(t_2) > Pr(t_1)$, then R would play a_2. But that would mean that $U^*_S(t_1) = 0$, while $U^*_S(t_2) = 1$. In this scenario, message "I am of type t_1" is credible, under van Rooy's (2003) notion, but "I am of type t_2" is not, because it is not relevant. Notice that if a speaker is of type t_2, he wouldn't say anything, but the fact that the speaker didn't say anything, if taken into account, must be interpreted as S being of type t_2 (because otherwise S would have said 'I am t_1'.) Assuming that saying nothing is saying the trivial proposition, R can conclude something more from some messages than is literally expressed. This is not unlike conversational implicatures (Grice, 1975).

So far we have seen that if preferences are aligned, a notion of credibility helps predict successful communication in a natural way. What about circumstances where this ideal condition is not satisfied? Look at the following table:

(5)

$U(t_i, a_j)$	a_1	a_2
t_1	1,1	0,0
t_2	1,0	0,1

In this case, both types of S want R to play a_1 and R would do so, in case he believed that S is of type t_1. However, R will not believe S's message "I am of type t_1", because if S is of type t_2 she still wants R to believe that she is of type t_1, and thus wants to mislead the receiver. Credible communication is not possible now. More in general, it can be shown that costless messages with a pre-existing meaning can be used to credibly transmit information only if it is known by the receiver that it is in the sender's interest to speak the truth.[6] If communicative manipulation is predicted to be possible at all, its successful use is predicted to be highly restricted.

We also must acknowledge that a proper notion of messages credibility is more complicated that indicated so far. Essentially, Farrell's notion and the slight amendment

[5] Benz (2007) criticizes this and other decision-theoretic approaches, arguing for the need to take the speaker's perspective into account (c.f. Benz, 2006; Benz and van Rooij, 2007, for models where this is done). In particular, Benz; Benz proved that any speaker strategy aiming at the maximization of relevance necessarily produces misleading utterances. This, according to Benz, entails that relevance maximization alone is not sufficient to guarantee credibility.

[6] The most relevant game-theoretical contributions are by Farrell (1988, 1993), Rabin (1990), Matthews et al. (1991), Zapater (1997). More recently, this topic has been reconsidered from a more linguistic point of view, e.g., by Stalnaker (2006), Franke (2010) and Franke et al. (2012).

we introduced above use a *forward induction* argument to show that agents can talk themselves out of an equilibrium. But it seems we didn't go far enough. To show this, consider the following game where states are again assumed equiprobable:

(6)

$U(t_i, a_j)$	a_1	a_2	a_3	a_4
t_1	10,5	0,0	1,4.1	-1,3
t_2	0,0	10,5	1,4.1	-1,3
t_3	0,0	0,0	1,4.1	-1,6

Let's suppose again that we start with a situation with only the decision problem and no communication. In this case, R responds with a_3. According to Farrell, this gives rise to two credible announcements: "I am of type t_1" and "I am of type t_2", with the obvious best responses. This is because both types t_1 and t_2 can profit from having these true messages believed: a credulous receiver will answer with actions a_1 and a_2 respectively. A speaker of type t_3 cannot make a credible statement, because revealing her identity would only lead to a payoff strictly worse than what she obtains if R plays a_3. Consequently, R should respond to no message with the same action as he did before, i.e., a_3. But once R realizes that S could have made the other statements credibly, but didn't, she will realize that the speaker must have been of type t_3 and will respond with a_4, and not with a_3. What this shows is that to account for the credibility of a message, one needs to think of higher levels of strategic sophistication. This also suggests that if either R or S do not believe in common belief in rationality, then misleading communication might again be possible. This is indeed what we will come back to presently in Section 1.3.

But before turning to that, we should address one more general case. Suppose we assume that messages not only have a semantic meaning, but that speakers also obey Grice's Maxim of Quality and do not assert falsehoods (Grice, 1975).[7] Do we predict more communication now? Milgrom and Roberts (1986) demonstrate that in such cases it is best for the decision maker to "assume the worst" about what S reports and that S has omitted information that would be useful. Milgrom and Roberts show that the optimal equilibrium strategy will always be the *sceptical posture*. In this situation, S will know that, unless the decision maker is told everything, the decision maker will take a stance against both his own interests (had he had full information) and the interests of S. Given this, the S could as well reveal all she knows.[8] This means that when speakers might try to manipulate the beliefs of the decision maker by being less precise than they could be, this won't help because an ideally rational decision maker will see through this attempt of manipulation. In conclusion, manipulation by communication is impossible in this situation; a result that is very much in conflict with what we perceive daily.[9]

[7] It is very frequently assumed in game theoretic models of pragmatic reasoning that the sender is compelled to truthful signaling by the game model. This assumption is present, for instance, in the work of (Parikh, 1991, 2001, 2010), but also assumed by many others. As long as interlocutors are cooperative in the Gricean sense, this assumption might be innocuous enough, but, as the present considerations make clear, are too crude a simplification when we allow conflicts of interest.

[8] The argument used to prove the result is normally called the *unraveling argument*. See Franke et al. (2012) for a slightly different version.

[9] Shin (1994) proves a generalization of Milgrom and Roberts's (1986) result, claiming that there always

1.3 Manipulation & bounded rationality

Many popular and successful theories of meaning and communicative behaviour are based on theories of ideal reasoning and rational behavior. But there is a lot of theoretical and experimental evidence that human beings are not perfectly rational reasoners. Against the assumed idealism it is often held, for instance, that we sometimes hold inconsistent beliefs, and that our decision making exhibits systematic biases that are unexplained by the standard theory (e.g. Simon, 1959; Tversky and Kahnemann, 1974). From this point of view, standard game theory is arguably based on a number of unrealistic assumptions. We will address two of such assumptions below, and indicate what might result if we give these up. First we will discuss the assumption that the game being played is common knowledge. Then we will investigate the implications of giving up the hypothesis that everybody is ideally rational, and that this is common knowledge. Finally, we will discuss what happens if our choices are systematically biased. In all three cases, we will see more room for successful manipulation.

Unawareness of the game being played. In standard game theory it is usually assumed that players conceptualize the game in the same way, i.e., that it is common knowledge what game is played. But this seems like a highly idealized assumption. It is certainly the case that interlocutors occasionally operate under quite different conceptions of the context of conversation, i.e., the 'language game' they are playing. This is evidenced by misunderstandings, but also by the way we talk: cooperative speakers must not only provide information but also enough background to make clear how that information is relevant. To cater for these aspects of conversation, Franke (forthcoming) uses models for *games with unawareness* (c.f. Halpern and Rêgo, 2006; Feinberg, 2011a; Heifetz et al., 2012) to give a general model for pragmatic reasoning in situations where interlocutors may have variously diverging conceptualizations of the context of utterance relevant to the interpretation of an utterance, different beliefs about these conceptualizations, different beliefs about these beliefs and so on. However, Franke (forthcoming) only discusses examples where interlocutors are well-behaved Gricean cooperators (Grice, 1975) with perfectly aligned interests. Looking at cases where this is not so, Feinberg (2008, 2011b) demonstrates that taking unawareness into account also provides a new rationale for communication in case of conflicting interests. Feinberg gives examples where communicating one's awareness of the set of actions which the decision maker can choose from might be beneficial for both parties involved. But many other examples exist (e.g. Ozbay, 2007). Here is a very simple one that nonetheless demonstrates the relevant conceptual points.

Reconsider the basic case in (5) that we looked at previously. We have two types of senders: t_1 wants his type to be revealed, and t_2 wishes to be mistaken for a type t_1. As we saw above, the message "I am of type t_1" is not credible in this case, because a sender of type t_2 would send it too. Hence, a rational decision maker should not believe that the actual type is t_1 when he hears that message. But if the decision maker is not aware that there could be a type t_2 that might want to mislead him, then, although

exists a sequential equilibrium (a strengthened notion of Nash equilibrium we have not introduced here) of the persuasion game in which the sender's strategy is perfectly revealing in the sense that the sender will say exactly what he knows.

incredible from the point of view of a perfectly aware spectator, from the decision maker's subjective point of view, the message "I'm of type t_1" is perfectly credible. The example is (almost) entirely trivial, but the essential point nonetheless significant. If we want to mislead, but also if we want to reliably and honestly communicate, it might be the very best thing to do to leave the decision maker completely in the dark as to any mischievous motivation we might pursue or, contrary to fact, might have been pursuing.

This simple example also shows the importance of choosing, not only *what* to say, but also *how* to say it. (We will come back to this issue in more depth below when we look at *framing effects*.) In the context of only two possible states, the messages "I am of type t_1" and "I am not of type t_2" are equivalent. But, of course, from a persuasion perspective they are not equally good choices. The latter would make the decision maker aware of the type t_2, the former need not. So although contextually equivalent in terms of their extension, the requirements of efficient manipulation clearly favor the one over the other simply in terms of surface form, due to their variable effects on the awareness of the decision maker.

In a similar spirit, van Rooij and Franke (2012) use differences in awareness-raising of otherwise equivalent conditionals and disjunctions to explain why there are conditional threats (7a) and promises (7b), and also disjunctive threats (7c), but, what is surprising from a logical point of view, no disjunctive promises (7d).

(7) a. If you don't give me your wallet, I'll punish you severely. threat
b. If you give me your wallet, I'll reward you splendidly. promise
c. You will give me your wallet or I'll punish you severely. threat
d. ? You will not give me your wallet or I'll reward you splendidly. threat

Sentence (7d) is most naturally read as a threat by accommodating the admittedly aberrant idea that the hearer has a strong aversion against a splendid reward. If that much accommodation is impossible, the sentence is simply pragmatically odd. The general absence of disjunctive threats like (7d) from natural language can be explained, van Rooij and Franke argue, by noting that these are suboptimal manipulation strategies because, among other things, they raise the possibility that the speaker does *not* want the hearer to perform. Although conditional threats also might make the decision maker aware of the "wrong" option, these can still be efficient inducements because, according to van Rooij and Franke (2012) the speaker can safely increase the stakes, by committing to more severe levels of punishment. If the speaker would do that for disjunctive promises, she would basically harm herself by expensive promises.

These are just a few basic examples that show how reasoning about the possibility of subjective misconceptions of the context/game model affects what counts as an optimal manipulative technique. But limited awareness of the context model is not the only cognitive limitation that real life manipulators may wish to take into consideration. Limited reasoning capacity is another.

No common knowledge of rationality. A number of games can be solved by (iterated) elimination of dominated strategies. If we end up with exactly one (rationalizable) strategy for each player, this strategy combination must be a Nash equilibrium.

Even though this procedure seems very appealing, it crucially depends on a very strong epistemic assumption: *common knowledge of rationality*; not only must every agent be ideally rational, everybody must also know of each other that they are rational, and they must know that they know it, and so on *ad infinitum*.[10] However, there exists a large body of empirical evidence that the assumption of common knowledge of rationality is highly unrealistic (c.f. Camerer, 2003, Chapter 5). Is it possible to explain deception and manipulation if we give up this assumption?

Indeed, it can be argued that whenever we do see attempted deceit in real life we are sure to find at least a belief of the deceiver (whether justified or not) that the agent to be deceived has some sort of limited reasoning power that makes the deception at least conceivably successful. Some agents are more sophisticated than others, and think further ahead. To model this, one can distinguish different *strategic types* of players, often also referred to as *cognitive hierarchy models* within the economics literature (e.g. Camerer et al., 2004; Rogers et al., 2009) or as *iterated best response models* in game theoretic pragmatics (e.g. Jäger, 2011; Jäger and Ebert, 2009; Franke, 2011). A strategic type captures the level of strategic sophistication of a player and corresponds to the number of steps that the agent will compute in a sequence of iterated best responses. One can start with an unstrategic level-0 players. An unstrategic level-0 hearer (a credulous hearer), for example, takes the semantic content of the message he receives literally, and doesn't think about why a speaker used this message. Obviously, such a level-0 receiver can sometimes be manipulated by a level-1 sender. But such a sender can in turn be outsmarted by a level-2 receiver, etc. In general, a level-$(k+1)$ player is one who plays a best response to the behavior of a level-k player. (A *best response* is a rationally best reaction to a given belief about the behavior of all other players.) A fully sophisticated agent is a level-ω player who behaves rationally given her belief in common belief in rationality.

Using such cognitive hierarchy models, Crawford (2003), for instance, showed that in case sender and/or receiver believe that there is a possibility that the other player is less sophisticated than he is himself, deception is possible (c.f. Crawford, 2007). Moreover, even sophisticated level-ω players can be deceived if they are not sure that their opponents are level-ω players too. Crawford assumed that messages have a specific semantic content, but did not presuppose that speakers can only say something that is true.

Building on work of Rabin (1990) and Stalnaker (2006), Franke (2010) offers a notion of *message credibility* in terms of an iterated best response model (see also Franke, 2009, Chapter 2). The general idea is that the conventional meaning of a message is a strategically non-binding *focal point* that defines the behavior of unstrategic level-0 players. For instance, for the simple game in (5), a level-0 receiver would be credulous and believe that message "I am of type t_2" is true and honest. But then a level-1 sender of type t_2 would exploit this naïve belief and also believe that her deceit is successful. Only if the receiver in fact is more sophisticated than that, would he see through the deception. Roughly speaking, a message is then considered credible iff no strategic sender type would ever like to use it falsely. In effect, this model not only provably

[10] We are rather crudely glossing here over many interesting subtleties in the notion of rationality and (common) belief in it. See, for instance, the contributions by Bonnano, Pacuit and Perea to this volume.

improves on the notion of message credibility, but also explains when deceit can be (believed to be) successful.

We can conclude that (i) it might be unnatural to assume common knowledge of rationality, and (ii) by giving up this assumption, we can explain much better why people communicate than standard game theory can: sometimes we communicate to manipulate others on the assumption that the others don't see it through, i.e., that we are smarter than them (whether this is justified or not).

Framing. As noted earlier, there exists a lot of experimental and theoretical evidence that we do not, and even cannot, always pick our choices in the way we should do according to the standard normative theory. In decision theory it is standardly assumed, for instance, that preference orders are transitive and complete. Still, already May (1945) has shown that cyclic preferences were not extraordinary (violating transitivity of the preference relation), and Luce (1959) noted that people sometimes seem to choose one alternative over another with a given consistent probability not equal to one (violating completeness of the preference relation). What is interesting for us is that due to the fact that people don't behave as rationally as the standard normative theory prescribes, it becomes possible for smart communicators to *manipulate* them: to convince them to do something that goes against their own interest. We mentioned already the use of *false implication*. Perhaps better known is the *money pump* argument: the fact that agents with intransitive preferences can be exploited because they are willing to participate in a series of bets where they will loose for sure. Similarly, manipulators make use of *false analogies*. According to psychologists, reasoning by analogy is used by boundedly rational agents like us to reduce the evaluation of new situations by comparing them with familiar ones (c.f. Gilboa and Schmeidler, 2001). Though normally a useful strategy, it can be exploited. There are many examples of this. Just to take one, in an advertisement for Chanel No. 5, a bottle of the perfume is pictured together with Nicole Kidman. The idea is that Kidman's glamour and beauty is transferred from her to the product. But perhaps the most common way to influence a decision maker making use of the fact that he or she does not choose in the prescribed way is by *framing*.

By necessity, a decision maker interprets her decision problem in a particular way. A different interpretation of the same problem may sometimes lead to a different decision. Indeed, there exists a lot of experimental evidence, that our decision making can depend a lot on how the problem is set. In standard decision theory it is assumed that decisions are made on the basis of information, and that it doesn't matter how this information is presented. It is predicted, for instance, that it doesn't matter whether you present this glass as being half full, or as half empty. The fact that it sometimes does matter is called the *framing effect*. This effect can be used by manipulators to present information such as to influence the decision maker in their own advantage. An agent's choice can be manipulated, for instance, by the addition or deletion of other 'irrelevant' alternative action to choose between, or by presenting the action the manipulator wants to be chosen in the beginning of, or at multiple times in, the set of alternative actions.

Framing is possible, because we apparently do not always choose by maximizing utility. Choosing by maximizing expected utility, the decision maker integrates the

expected utility of an action with what he already has. Thinking for simplicity of utility just in terms of monetary value, it is thus predicted that someone who starts with 100 Euros and gains 50, ends up being equally happy as one who started out with 200 Euros and lost 50. This prediction is obviously wrong, and the absurdity of the prediction was highlighted especially by Kahneman and Tversky. They pointed out that decision makers think in terms of *gains* and *losses* with respect to a *reference point*, rather than in terms of context-independent utilities as the standard theory assumes. This reference point typically represents what the decision maker currently has, but—and crucial for persuasion—it need not be. Another, in retrospect, obvious failure of the normative theory is that they systematically overestimate low-probability events. How else can one explain why people buy lottery tickets and pay quite some money to insure themselves against very unlikely losses?

Kahneman and Tversky brought to light less obvious violations of the normative theory as well. Structured after the well-known Allais paradox, their famous Asian disease experiment (Tversky and Kahnemann, 1981), for instance, shows that in most people's eyes, a *sure* gain is worth more than a *probable* gain with an equal or greater expected value. Other experiments by the same authors show that the opposite is true for losses. People tend to be risk-averse in the domain of gains, and risk-taking in the domain of losses, where the displeasure associated with the loss is greater than the pleasure associated with the same amount of gains.

Notice that as a result, choices can depend on whether outcomes are seen as gains or losses. But whether something is seen as a gain or a loss depends on the chosen reference-point. What this reference-point is, however, can be influenced by the manipulator. If you want to persuade parents to vaccinate their children, for instance, one can set the outcomes either as losses, or as gains. Experimental results show that persuasion is more successful by loss-framed than by gain-framed appeals (O'Keefe and Jensen, 2007).

Framing effects are predicted by Kahneman and Tversky's *Prospect Theory*: a theory that implements the idea that our behavior is only boundedly rational. But if correct, it is this kind of theory that should be taken into account in any serious analysis of persuasive language use.

Summary. Under idealized assumptions about agents' rationality and knowledge of the communicative situation, manipulation by strategic communication is by and large impossible. Listeners see through attempts of deception and speakers therefore do not even attempt to mislead. But manipulation can prosper among boundedly rational agents. If the decision maker is unaware of some crucial parts of the communicative situation (most palpably: the mischievous intentions of the speaker) or if the decision maker does not apply strategic reasoning deeply enough, deception may be possible. Also if the manipulator, but not the decision maker, is aware of the cognitive biases that affect our decision making, these mechanism can be exploited as well.

2 Opinion dynamics & efficient propaganda

While the previous section focused exclusively on the pragmatic dimension of persuasion, investigating *what* to say and *how* to say it, there is a wider social dimension to successful manipulation as well: determining *who we should address*. In this section, we will assume that agents are all part of a social network, and we will discuss how to best propagate one's own ideas through a social network.

We present a novel variant of DeGroot's classical model of opinion dynamics (DeGroot, 1974) that allows us to address the question how an agent, given his position in a social web of influenceability, should try to strategically influence others, so as to maximally promote her opinion in the relevant population. More concretely, while DeGroot's model implicitly assumes that agents distribute their persuasion efforts equally among the neighbors in their social network, we consider a new variant of DeGroot's model where a small fraction of players is able to re-distribute their persuasion efforts strategically. Using numerical simulations, we try to chart the terrain of more or less efficient opinion-promoting strategies and conclude that in order to successfully promote your opinion in your social network you should: (i) spread your web of influence wide (i.e., not focussing all effort on a single or few individuals), (ii) choose "easy targets" for quick success and "influential targets" for long-term success, and (iii), if possible, coordinate your efforts with other influencers so as to get out of each other's way. Which strategy works best, however, depends on the interaction structure of the population in question. The upshot of this discussion is that, even if computing the theoretically optimal strategy is out of the question for a resource-limited agent, the more an agent can exploit rudimentary or even detailed knowledge of the social structure of a population, the better she will be able to propagate her opinion.

Starting Point: The DeGroot Model. DeGroot (1974) introduced a simple model of opinion dynamics to study under which conditions a consensus can be reached among all members of a society (cf. Lehrer, 1975). DeGroot's classical model is a round-based, discrete and linear update model.[11] Opinions are considered at discrete time steps $t \in \mathbb{N}^{\geq 0}$. In the simplest case, an opinion is just a real number, representing, e.g., to what extent an agent endorses a position. For n agents in the society we consider the row vector of opinions $\mathbf{x}(t)$ with $\mathbf{x}(t)^T = \langle x_1(t), \ldots, x_n(t) \rangle \in \mathbb{R}^n$ where $x_i(t)$ is the opinion of agent i at time t.[12] Each round all agents update their opinions to a weighted average of the opinions around them. Who influences whom how much is captured by *influence matrix* P, which is a (row) stochastic $n \times n$ matrix with P_{ij} the weight with which agent i takes agent j's opinion into account. DeGroot's model then considers the simple linear update in (1):[13]

$$\mathbf{x}(t+1) = P\,\mathbf{x}(t)\,. \tag{1}$$

[11] DeGroot's model can be considered as a simple case of Axelrod's (1997) famous model of cultural dynamics (c.f. Castellano et al., 2009, for overview).

[12] We write out that transpose $\mathbf{x}(t)^T$ of the row vector $\mathbf{x}(t)$, so as not to have to write its elements vertically.

[13] Recall that if A and B are (n, m) and (m, p) matrices respectively, then $A\,B$ is the matrix product with $(A\,B)_{ij} = \sum_{k=1}^{m} A_{ik} B_{ki}$.

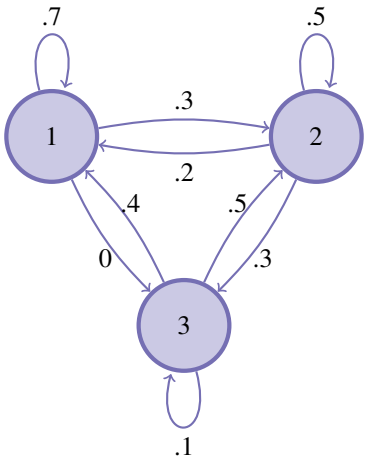


Figure 1: Influence in a society represented as a (fully connected, weighted and directed) graph.

For illustration, suppose that the society consists of just three agents and that influences among these are given by:

$$P = \begin{pmatrix} .7 & .3 & 0 \\ .2 & .5 & .3 \\ .4 & .5 & .1 \end{pmatrix}. \tag{2}$$

The rows in this influence matrix give the proportions with which each agent updates her opinions at each time step. For instance, agent 3's opinion at time $t+1$ is obtained by taking .4 parts of agent 1's opinion at time t, .5 parts of agent 2's and .1 parts of her own opinion at time t. For instance, if the vector of opinions at time $t = 0$ is a randomly chosen $\mathbf{x}(0)^T = \langle .6, .2, .9 \rangle$, then agent 3's opinion at the next time step will be $.4 \times .6 + .5 \times .2 + .1 \times .9 \approx .43$. By equation (1), we compute these updates in parallel for each agent, so we obtain $\mathbf{x}(1)^T \approx \langle .48, .49, .43 \rangle$, $\mathbf{x}(2)^T \approx \langle .48, .47, .48 \rangle$ and so on.[14]

DeGroot's model acknowledges the social structure of the society of agents in its specification of the influence matrix P. For instance, if $p_{ij} = 0$, then agent i does not take agent j's opinion into account at all; if $p_{ii} = 1$, then agent i does not take anyone else's opinion into account; if $p_{ij} < p_{ik}$, then agent k has more influence on the opinion of agent i than agent j.

It is convenient to think of P as the adjacency matrix of a fully-connected, weighted and directed graph, as shown in Figure 1. As usual, rows specify the weights of outgo-

[14] In this particular case, opinions converge to a consensus where everybody holds the same opinion. In his original paper DeGroot showed that, no matter what $\mathbf{x}(0)$, if P has at least one column with only positive values, then, as t goes to infinity, $\mathbf{x}(t)$ converges to a unique vector of uniform opinions, i.e., the same value for all $\mathbf{x}_i(t)$. Much subsequent research has been dedicated to finding sufficient (and necessary) conditions for opinions to converge or even to converge to a consensus (c.f. Jackson, 2008; Acemoglu and Ozdaglar, 2011, for overview). Our emphasis, however, will be different, so that we sidestep these issues.

ing connections, so that we need to think of a weighted edge in a graph like in Figure 1 as a specification of how much an agent (represented by a node) "cares about" or "listens to" another agent's opinion. The agents who agent i listens to, in this sense, are the *influences* of i:

$$I(i) = \left\{ j \mid p_{ij} > 0 \wedge i \neq j \right\} .$$

Inversely, let's call all those agents that listen to agent i as the *audience* of i:

$$A(i) = \left\{ j \mid p_{ji} > 0 \wedge i \neq j \right\} .$$

One more notion that will be important later should be mentioned here already. Some agents might listen more to themselves than others. Since how much agent i holds on to her own opinion at each time step is given by value p_{ii}, the diagonal $\texttt{diag}(P)$ of P can be interpreted as the vector of the agents' *stubbornness*. For instance, in example (2) agent 1 is the most stubborn and agent 3 the least convinced of his own views, so to speak.

Strategic Promotion of Opinions. DeGroot's model is a very simple model of how opinions might spread in a society: each round each agent simply adopts the weighted average of the opinions of his influences, where the weights are given by the fixed influence matrix. More general update rules than (1) have been studied, e.g., ones that make the influence matrix dependent on time and/or the opinions held by other agents, so that we would define $\mathbf{x}(t+1) = P(t, \mathbf{x}(t))\, \mathbf{x}(t)$ (cf. Hegselmann and Krause, 2002). We are interested here in an even more liberal variation of DeGroot's model in which (some of the) agents can *strategically* determine their influence, so as to best promote their own opinion. In other terms, we are interested in opinion dynamics of the form:

$$\mathbf{x}(t+1) = P(S)\, \mathbf{x}(t) , \tag{3}$$

where P depends on an $n \times n$ *strategy matrix* S where each row S_i is a strategy of agent i and each entry S_{ij} specifies how much effort agent i invests in trying to inpose her current opinion on each agent j.

Eventually we are interested in the question when S_i is a *good* strategy for a given influence matrix P, given that agent i wants to promote her opinion as much as possible in the society. But to formulate and address this question more precisely, we first must define (i) what kind of object a strategy is in this setting and (ii) how exactly the *actual influence matrix* $P(S)$ is computed from a given strategy S and a given influence matrix P.

Strategies. We will be rather liberal as to how agents can form their strategies: S could itself depend on time, the current opinions of others etc. We will, however, impose two general constraints on S because we want to think of *strategies as allocations of persuasion effort*. The first constraint is a mere technicality, requiring that $S_{ii} = 0$ for all i: agents do not invest effort into manipulating themselves. The second constraint is that each row vector S_i is a stochastic vector, i.e., $S_{ij} \geq 0$ for all i and j and $\sum_{j=1}^{n} S_{ij} = 1$ for all i. This is to make sure that strengthening one's influence on some

agents comes at the expense of weakening one's influence on others. Otherwise there would be no interesting strategic considerations as to where best to exert influence. We say that S_i is a *neutral strategy* for P if it places equal weight on all j that i can influence, i.e., all $j \in A(i)$.[15] We call S neutral for P, if S consists entirely of neutral strategies for P. We write S^* for the neutral strategy of an implicitly given P.

Examples of strategy matrices for the influence matrix in (2) are:

$$S = \begin{pmatrix} 0 & .9 & .1 \\ .4 & 0 & .6 \\ .5 & .5 & 0 \end{pmatrix} \qquad S' = \begin{pmatrix} 0 & .1 & .9 \\ .5 & 0 & .5 \\ 0 & 1 & 0 \end{pmatrix} \qquad S^* = \begin{pmatrix} 0 & .5 & .5 \\ .5 & 0 & .5 \\ 0 & 1 & 0 \end{pmatrix}.$$

According to strategy matrix S, agent 1 places .9 parts of her available persuasion effort on agent 2, and .1 on agent 3. Notice that since in our example in (2) we had $P_{13} = 0$, agent 3 cannot influence agent 1. Still, nothing prevents her from allocating persuasion effort to agent 1. (This would, in a sense, be irrational but technically possible.) That also means that S_3 is *not* the neutral strategy for agent 3. The neutral strategy for agent 3 is S'_3 where all effort is allocated to the single member in agent 3's audience, namely agent 2. Matrix S' also includes the neutral strategy for agent 2, who has two members in her audience. However, since agent 1 does not play a neutral strategy in S', S' is not neutral for P, but S^* is.

Actual Influence. Intuitively speaking, we want the actual influence matrix $P(S)$ to be derived by adjusting the influence weights in P by the allocations of effort given in S. There are many ways in which this could be achieved. Our present approach is motivated by the desire to maintain a tight connection with the original DeGroot model. We would like to think of (1) as the special case of (3) where every agent plays a neutral strategy. Concretely, we require that $P(S^*) = P$. (Remember that S^* is the neutral strategy for P.) This way, we can think of DeGroot's classical model as a description of opinion dynamics in which no agent is a strategic manipulator, in the sense that no agent deliberately tries to spread her opinion by exerting more influence on some agents than on others.

We will make one more assumption about the operation $P(S)$, which we feel is quite natural, and that is that $\texttt{diag}(P(S)) = \texttt{diag}(P)$, i.e., the agents' stubbornness should not depend on how much they or anyone else allocates persuasion effort. In other words, strategies should compete only for the resources of opinion change that are left after subtracting an agent's stubbornness.

To accommodate these two requirements in a natural way, we define $P(S)$ with respect to a reference point formed by the neutral strategy S^*. For any given strategy matrix S, let $\overline{S}$ be the column-normalized matrix derived from S. $\overline{S}_{ij}$ is i's *relative persuasion effort* affecting j, when taking into account how much everybody invests in influencing j. We compare $\overline{S}$ to the relative persuasion effort $\overline{S^*}$ under the neutral strategy: call $R = \overline{S}/\overline{S^*}$ the matrix of *relative net influences* given strategy S.[16] The actual influence matrix $P(S) = Q$ is then defined as a reweighing of P by the relative

[15] We assume throughout that $A(i)$ is never empty.

[16] Here and in the following, we adopt the convention that $x/0 = 0$.

net influences R:

$$Q_{ij} = \begin{cases} P_{ij} & \text{if } i = j \\ \frac{P_{ij}R_{ji}}{\sum_k P_{ik}R_{ki}}(1 - P_{ii}) & \text{otherwise.} \end{cases} \tag{4}$$

Here is an example illustrating the computation of actual influences. For influence matrix P and strategy matrix S we get the actual influences $P(S)$ as follows:

$$P = \begin{pmatrix} 1 & 0 & 0 \\ .2 & .5 & .3 \\ .4 & .5 & .1 \end{pmatrix} \qquad S = \begin{pmatrix} 0 & .9 & .1 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \qquad P(S) \approx \begin{pmatrix} 1 & 0 & 0 \\ .27 & .5 & .23 \\ .12 & .78 & .1 \end{pmatrix}.$$

To get there we need to look at the matrix of relative persuasion effort $\overline{S}$ given by S, the neutral strategy S^* for this P and the relative persuasion effort $\overline{S^*}$ under the neutral strategy:

$$\overline{S} = \begin{pmatrix} 0 & 9/19 & 1/11 \\ 0 & 0 & 10/11 \\ 0 & 10/19 & 0 \end{pmatrix} \qquad S^* = \begin{pmatrix} 0 & .5 & .5 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \qquad \overline{S^*} = \begin{pmatrix} 0 & 1/3 & 1/3 \\ 0 & 0 & 2/3 \\ 0 & 2/3 & 0 \end{pmatrix}.$$

That $\overline{S^*}_{12} = 1/3$, for example, tells us that agent 1's influence on agent 2 $P_{21} = 1/5$ comes about in the neutral case where agent 1 invests half as much effort into influencing agent 2 as agent 3 does. To see what happens when agent 1 plays a non-neutral strategy, we need to look at the matrix of relative net influences $R = \overline{S}/\overline{S^*}$, which, intuitively speaking, captures how much the actual case $\overline{S}$ deviates from the neutral case $\overline{S^*}$:

$$R = \begin{pmatrix} 0 & 27/19 & 3/11 \\ 0 & 0 & 15/11 \\ 0 & 15/19 & 0 \end{pmatrix}.$$

This derives $P(S) = Q$ by equation (4). We spell out only one of four non-trivial cases here:

$$\begin{aligned} Q_{21} &= \frac{P_{21}R_{12}}{P_{11}R_{11} + P_{12}R_{21} + P_{13}R_{31}}(1 - P_{22}) \\ &= \frac{2/10 \times 27/19}{1/5 \times 27/19 + 1/2 \times 0 + 3/10 \times 15/19}(1 - 1/2) \\ &\approx 0.27 \end{aligned}$$

In words, by investing 9 times as much into influencing agent 2 than into influencing agent 3, agent 1 gains effective influence of ca. $.27 - .2 = .07$ over agent 2, as compared to when she neutrally divides effort equally among her audience. At the same time, agent 1 looses effective influence of ca. $.4 - .12 = .28$ on agent 3. (This strategy might thus seem to only diminish agent 1's actual influence in the updating process. But, as we will see later on, this can still be (close to) the optimal choice in some situations.)

It remains to check that the definition in (4) indeed yields a conservative extension of the classical DeGroot-process in (1):

Fact 1. $P(\overline{S}) = P$.

Proof. Let $Q = P(\overline{S})$. Look at arbitrary Q_{ij}. If $i = j$, then trivially $Q_{ij} = P_{ij}$. If $i \neq j$, then

$$Q_{ij} = \frac{P_{ij}R_{ji}}{\sum_k P_{ik}R_{ki}}(1 - P_{ii}),$$

with $R = \overline{S^*}/\overline{S^*}$. As $S_{ii} = 0$ by definition of a strategy, we also have $R_{ii} = 0$. So we get:

$$Q_{ij} = \frac{P_{ij}R_{ji}}{\sum_{k\neq i} P_{ik}R_{ki}}(1 - P_{ii}).$$

Moreover, for every $k \neq i$, $R_{kl} = 1$ whenever $P_{lk} > 0$, otherwise $R_{kl} = 0$. Therefore:

$$Q_{ij} = \frac{P_{ij}}{\sum_{k\neq i} P_{ik}}(1 - P_{ii}) = P_{ij}.$$

□

The Propaganda Problem. The main question we are interested in is a very general one:

(8) *Propaganda problem (full)*: Which individual strategies S_i are good or even optimal for promoting agent i's opinion in society?

This is a game problem because what is a good promotion strategy for agent i depends on what strategy all other agents play as well. As will become clear below, the complexity of the full propaganda problem is daunting. We therefore start first by asking a simpler question, namely:

(9) *Propaganda problem (restricted, preliminary)*: Supposing that most agents behave non-strategically like agents in DeGroot's original model (call them: *sheep*), which (uniform) strategy should a minority of strategic players (call them: *wolves*) adopt so as best to promote their minority opinion in the society?

In order to address this more specific question, we will assume that initially wolves and sheep have opposing opinions: if i is a wolf, then $x_i(0) = 1$; if i is a sheep, then $x_i(0) = -1$. We could think of this as being politically right wing or left wing; or of endorsing or rejecting a proposition, etc. Sheep play a neutral strategy and are susceptible to opinion change ($P_{ii} < 1$ for sheep i). Wolves are maximally stubborn ($P_{ii} = 1$ for wolves i) and can play various strategies. (For simplicity we will assume that all wolves in a population play the same strategy.) We are then interested in ranking wolf strategies with respect to how strongly they pull the community's *average opinion* $\bar{x}(t) = 1/n \times \sum_{i=1}^{n} x_i(t)$ towards the wolf opinion.

This formulation of the propaganda problem is still too vague to be of any use for categorizing good and bad strategies. We need to be more explicit at least about the number of rounds after which strategies are evaluated. Since we allow wolf strategies

to vary over time and/or to depend on other features which might themselves depend on time, it might be that some strategies are good at short intervals of time and others only after many more rounds of opinion updating. In other words, the version of the propaganda problem we are interested in here is dependent on the number of rounds k. For fixed P and $\mathbf{x}(0)$, say that $\mathbf{x}(k)$ results from a sequence of strategy matrices $\left\langle S^{(1)}, \ldots, S^{(k)} \right\rangle$ if for all $0 < i \leq k$: $\mathbf{x}(i) = P(S^{(i)})\,\mathbf{x}(i-1)$.

(10) *Propaganda problem (restricted, fixed P)*: For a fixed P, a fixed $\mathbf{x}(0)$ as described and a number of rounds $k > 0$, find a sequence of k strategy matrices $\left\langle S^{(1)}, \ldots, S^{(k)} \right\rangle$, with wolf and sheep strategies as described above, such that $\bar{\mathbf{x}}(k)$ is maximal for the $\mathbf{x}(k)$ that results from $\left\langle S^{(1)}, \ldots, S^{(k)} \right\rangle$.

What that means is that the notion of a *social influencing strategy* we are interested in here is that of an optimal *sequence* of k strategies, not necessarily a single strategy. Finding a good strategy in this sense can be computationally hard, as we would like to make clear in the following by a simple example. It is therefore that, after having established a feeling for how wolf strategies influence population dynamics over time, we will rethink our notion of a social influence strategy once more, arguing that the complexity of the problem calls for *heuristics* that are easy to apply yet yield good, if sub-optimal, results. But first things first.

Example: Lone-Wolf Propaganda. Although simpler than the full game problem, the problem formulated in (10) is still a very complex affair. To get acquainted with the complexity of the situation, let's look first at the simplest non-trivial case of a society of three agents with one wolf and two sheep: call it a *lone-wolf problem*. For concreteness, let's assume that the influence matrix is the one we considered previously, where agent 1 is the wolf:

$$P = \begin{pmatrix} 1 & 0 & 0 \\ .2 & .5 & .3 \\ .4 & .5 & .1 \end{pmatrix}. \tag{5}$$

Since sheep agents 2 and 3 are assumed to play a neutral strategy, the space of feasible strategies for this lone-wolf situation can be explored with a single parameter $a \in [0; 1]$:

$$S(a) = \begin{pmatrix} 0 & \text{a} & \text{1-a} \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}.$$

We can therefore calculate:

$$\overline{S^*} = \begin{pmatrix} 0 & 1/3 & 1/3 \\ 0 & 0 & 2/3 \\ 0 & 2/3 & 0 \end{pmatrix} \qquad \overline{S(a)} = \begin{pmatrix} 0 & a/a+1 & 1-a/2-a \\ 0 & 0 & 1/2-a \\ 0 & 1/a+1 & 0 \end{pmatrix}$$

$$R = \begin{pmatrix} 0 & 3a/a+1 & 3-3a/2-a \\ 0 & 0 & 3/4-2a \\ 0 & 3/2a+2 & 0 \end{pmatrix} \qquad P(S(a)) = \begin{pmatrix} 1 & 0 & 0 \\ 4a/8a+6 & 1/2 & 3/8a+6 \\ 36-36a/65-40a & 9/26-16a & 1/10 \end{pmatrix}$$

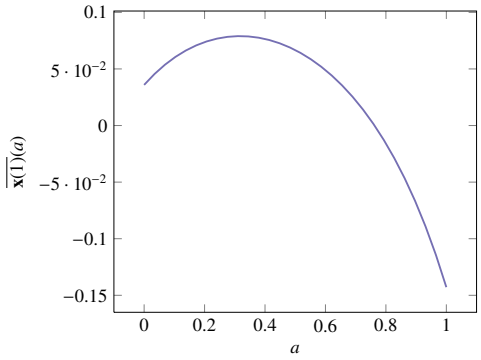


Figure 2: Population opinion after one round of updating with a strategy matrix $S(a)$ for all possible values of a, as described by the function in Equation (6).

Let's first look at the initial situation with $\mathbf{x}(0)^T = \langle 1, -1, -1 \rangle$, and ask what the best wolf strategy is for boosting the average population in just one time step $k = 1$. The relevant population opinion can be computed as a function of a, using basic algebra:

$$\overline{\mathbf{x}(1)}(a) = \frac{-224a^2 + 136a - 57}{-160a^2 + 140a + 195}\,. \tag{6}$$

This function is plotted in Figure 2. Another chunk of basic algebra reveals that this function has a local maximum at $a = .3175$ in the relevant interval $a \in [0; 1]$. In other words, the maximal shift towards wolf opinion in one step is obtained for the wolf strategy $\langle 0, .3175, .6825 \rangle$. This, then, is an exact solution to the special case of the propaganda problem state in (10) where P is given as above and $k = 1$.

How about values $k > 1$? Let's call any k-sequence of wolf strategies that maximizes the increase in average population opinion at each time step the `greedy` strategy. Notice that the `greedy` strategy does not necessarily select the same value of a in each round because each greedy choice of a depends on the actual sheep opinions x_2 and x_3. To illustrate this, Figure 3 shows (a numerical approximation of) the `greedy` values of a for the current example as a function of all possible sheep opinions. As is quite intuitive, the plot shows that the more, say, agent 3 already bears the wolf opinion, the better it is, when greedy, to focus persuasion effort on agent 2, and vice versa.

It may be tempting to hypothesize that strategy `greedy` solves the lone-wolf version of (10) for arbitrary k. But that's not so. From the fourth round onwards even playing the neutral strategy `sheep` (a constant choice of $a = 1/2$ in each round) is better than strategy `greedy`. This is shown in Figure 4, which plots the temporal development over 20 rounds of what we will call *relative opinion* for our current lone-wolf problem. Relative opinion of strategy X is the average population opinion as it develops under strategy X minus the average population opinion as it develops under baseline strategy `sheep`. Crucially, the plot shows that the relative opinion under `greedy` greedy choices falls below the baseline of non-strategic DeGroot play already very soon (after 3 rounds). This means that the influence matrix P we are looking at here provides

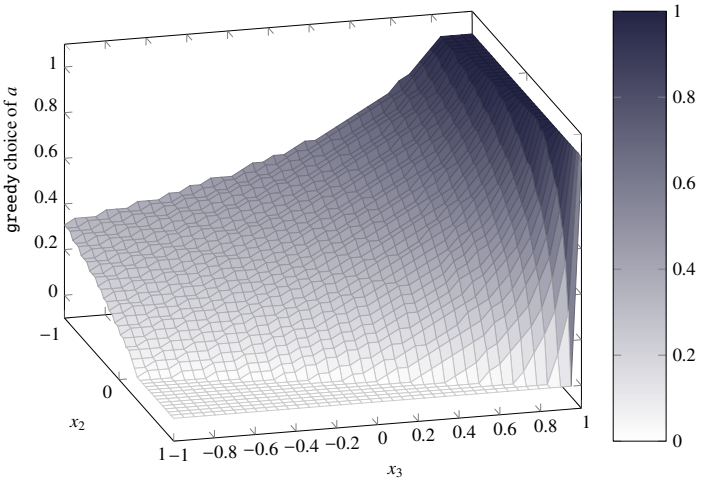


Figure 3: Dependency of the `greedy` strategy on the current sheep opinion for the lone-wolf problem given in (5). The graph plots the best choice of effort a to be allocated to persuading agent 2 for maximal increase of population opinion in one update step, as a function of all possible pairs of sheep opinions x_2 and x_3.

a counterexample against the *prima facie* plausible conjecture that playing `greedy` solves the propaganda problem in (10) for all k.

The need for heuristics. Of course, it is possible to calculate a sequence of a values for any given k and P that strictly maximizes the population opinion. But, as the previous small example should have made clear, the necessary computations are so complex that it would be impractical to do so frequently under "natural circumstances", such as under time pressure or in the light of uncertainty about P, the relevant k, the current opinions in the population etc. This holds in particular when we step beyond the lone-wolf version of the propaganda problem: with several wolves the optimization problem is to find the *set* of wolf strategies that are optimal *in unison*. Mathematically speaking, for each fixed P, this is a multi-variable, non-linear, constrained optimization problem. Oftentimes this will have a unique solution, but the computational complexity of the relevant optimization problem is immense. This suggests the usefulness, if not necessity of simpler, but still efficient *heuristics*.[17] For these reasons we focus in the following on intuitive and simple ways of playing the social manipulation game that make, for the most part, more innocuous assumptions about agents' computational capacities and knowledge of the social facts at hand. We try to demonstrate that these heuristics are not only simple, but also lead to quite good results on average, i.e., if

[17] Against this it could be argued that processes of evolution, learning and gradual optimization might have brought frequent manipulators at least close to the analytical optimum over time. But even then, it is dubious that the agents actually have the precise enough knowledge (of influence matrix P, current population opinion, etc.) to learn to approximate the optimal strategy. Due to reasons of learnability and generalizability, what evolves or is acquired and fine-tuned by experience, too, is more likely a good heuristic.

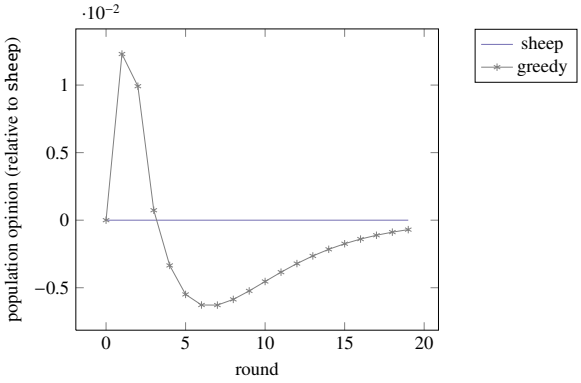


Figure 4: Temporal development of relative opinion (i.e., average population opinion relative to average population opinion under baseline strategy `sheep`) for several wolf strategies for the influence matrix in (5)

uniformly applied to a larger class of games.

To investigate the average impact of various strategies, we resort to numerical simulation. By generating many random influence matrices P and recording the temporal development of the population opinion under different strategies, we can compare the average success of these strategies against each other.

Towards efficient heuristics. For reasons of space, we will only look at a small sample of reasonably successful and resource efficient heuristics that also yield theoretical insights into the nature of the propaganda problem. But before going into details, a few general considerations about efficient manipulation of opinions are in order. We argue that in general for a manipulation strategy to be efficient it should: (i) not preach to the choir, (ii) target large groups, not small groups or individuals, (iii) take other manipulators into account, so as not to get into one another's way and (iv) take advantage of the social structure of society (as given by P). Let's look at all of these points in turn.

Firstly, it is obvious that any effort spent on a sheep which is already convinced, i.e., holds the wolf opinion one, is wasted.[18] A minimum standard for a rational wolf strategy would therefore be to spend no effort on audience members with opinion one as long as there are audience members with opinion lower than one. All of the strategies we look at below are implicitly assumed to conform to this requirement.

Secondly, we could make a distinction between strategies that place all effort onto just one audience member and strategies that place effort on more than one audience member (in the most extreme case that would be *all* of the non-convinced audience members). Numerical simulations show that, on average, strategies of the former kind clearly prove inferior to strategies of the latter kind. An intuitive argument why that is so is the following. For concreteness, consider the lone-wolf greedy maximization

[18]Strictly speaking, this can only happen in the limit, but this is an issue worth addressing, given (i) floating number imprecision in numerical simulations, and (ii) the general possibility (which we do not explicitly consider) of small independent fluctuations in agents' opinions.

problem plotted in Figure 2. (The argument holds in general.) Since the computation of $P(S)$ relies on the *relative* net influence R, playing extreme values ($a = 0$ or $a = 1$) is usually suboptimal because the influence gained on one agent is smaller than the influence lost on the other agent. This much concerns just one round of updating, but if we look at several rounds of updating, then influencing several agents to at least some extent is beneficial, because the increase in their opinion from previous rounds will lead to more steady increase in population opinion at later rounds too. All in all, it turns out that efficient manipulation of opinions, on a short, medium and long time scale, is achieved better if the web of influence is spread wide, i.e., if many or all suitable members of the wolves' audience are targeted with at least *some* persuasion effort. For simplicity, the strategies we consider here will therefore target all non-convinced members of each wolf's audience, but variably distribute persuasion effort among these.

Thirdly, another relevant distinction of wolf strategies is between those that are sensitive to the presence and behavior of other wolves and those that are not. The former may be expected to be more efficient, if implemented properly, but they are also more sophisticated. This is because they pose stronger requirements on the agents that implement these strategies: wolves who want to hunt in a pack should be aware of the other wolves and adapt their behavior to form an efficient *coalition strategy*. We will look at just one coalition strategy here, but find that, indeed, this strategy is (one of) the best from the small sample that is under scrutiny here. Surprisingly, the key to coalition success is not to join forces, but rather to get out of each other's way. Intuitively, this is because if several manipulators invest in influencing the same sheep, they thereby decrease their *relative* net influence unduly. On the other hand, if a group of wolves decides who is the main manipulator, then by purposefully investing little effort the other wolves boost the main manipulator's relative net influence.

Fourthly and finally, efficient opinion manipulation depends heavily on the social structure of the population, as given by P. We surely expect that a strategy which uses (approximate) knowledge of P in a smart way will be more effective than one that does not. The question is, of course, what features of the social structure to look at. Below we investigate two kinds of socially-aware heuristics: one that aims for sheep that can be easily influenced, and one that aims for sheep that are influential themselves. We expected that the former do better in the short run, while the latter might catch up after a while and eventually do better in the long run. This expectation is borne out, but exactly how successful a given strategy (type) is also depends on the structure of the society.

The cast. Next to strategy `sheep`, the strategies we look at here are called `influence`, `impact`, `eigenvector` and `communication`. We describe each in turn and then discuss their effectiveness, merits and weaknesses.

Strategy `influence` chooses a fixed value of a in every round, unlike the time-dependent `greedy`. Intuitively speaking, the strategy `influences` allocates effort among its audience proportional to how much influence the wolf has on each sheep: the more a member of an audience is susceptible to being influenced, the more effort is allocated to her. In effect, strategy `influence` says: "allocate effort relatively to how

much you are being listened to". In our running example with P as in Equation (5) the lone wolf has an influence on (sheep) agent 2 of $P_{12} = 1/5$ and of $P_{13} = 2/5$ on agent 3. Strategy `influence` therefore chooses $a = 1/3$, because the wolf's influence over agent 2 is half as big as that over agent 3.

Strategy `impact` is something like the opposite of strategy `influence`. Intuitively speaking, this strategy says: "allocate effort relatively to how much your audience is being listened to". The difference between `influence` and `impact` is thus that the former favors those the wolf has big influence over, while the latter favors those that have big influence themselves. To determine influence, strategy `impact` looks at the column vector P^T_j for each agent $j \in A(i)$ in wolf i's audience. This column vector P^T_j captures how much *direct influence* agent j has. We say that sheep j has more direct influence than sheep j' if the sum of the j-th column is bigger than that of the j'-th. (Notice that the rows, but not the columns of P must sum to one, so that some agents may have more direct influence than others.) If we look at the example matrix in equation (5), for instance, agent 2 has more direct influence than agent 3. The strategy `impact` then allocates persuasion effort proportional to relative direct influence among members of an audience. In the case at hand, this would lead to a choice of

$$a = \frac{\sum_k P_{k2}}{\sum_k P_{k2} + \sum_k P_{k3}} = 5/12\,.$$

Strategy `eigenvector` is very much like `impact`, but smarter, because it looks beyond *direct* influence. Strategy `eigenvector` for wolf i also looks at how influential the audience of members of i's audience is, how influential their audience is and so on *ad infinitum*. This transitive closure of social influence of all sheep can be computed with the (right-hand) eigenvector of the matrix P^*, where P^* is obtained by removing from P all rows and columns belonging to wolves.[19,20] For our present example, the right-hand unit eigenvector of matrix

$$P^* = \begin{pmatrix} .5 & .3 \\ .5 & .1 \end{pmatrix}$$

is approximately $\langle .679, .321 \rangle$. So the strategy `eigenvector` would choose a value of approximately $a = .679$ at each round.

Finally, we also looked at one coalition strategy, where wolves coordinate their behavior for better effect. Strategy `communication` is such a sophisticated coalition strategies that also integrates parts of the rationale behind strategy `influence`. Strategy `communication` works as follows. For a given target sheep i, we look at all wolves among the influences $I(i)$ of i. Each round a main manipulator is drawn from that group with a probability proportional to how much influence each potential manipulator has over i. Wolves then allocate 100 times more effort to each sheep in their audience for which they are the main manipulator in that round than to others. Since this much

[19] Removing wolves is necessary because wolves are the most influential players; in fact, since they are maximally stubborn, sheep would normally otherwise have zero influence under this measure.

[20] The DeGroot-process thereby gives a motivation for measures of eigenvector centrality, and related concepts such as the Google page-rank (cf. Jackson, 2008). Unfortunately, the details of this fascinating issue are off-topic in this context.

time-variable coordination seems only plausible, when wolves can negotiating their strategies each round, we refer to this strategy as `communication`.

We expect strategy `influence` and `communication` to have similar temporal properties, namely to outperform baseline strategy `sheep` in early rounds of play. `Communication` is expected to be better than `influence` because it is the more sophisticated coalition strategy. On the other hand, strategies `impact` and `eigenvector` should be better at later rounds of updating because they invest in manipulating influential or "central" agents of the society, which may be costly at first, but should pay off later on. We expect `eigenvector` to be better than `impact` because it is the more sophisticated social strategy that looks beyond *direct* influence at the *global* influence that agents have in the society.

Experimental set-up. We tested these predictions by numerical simulation in two experiments, each of which assumed a different *interaction structure* of the society of agents. The first experiment basically assumed that the society is *homogeneous*, in the sense that (almost) every wolf can influence (almost) every sheep and (almost) every sheep interacts with (almost) every sheep. The second experiment assumed that the pattern of interaction is *heterogeneous*, in the sense that who listens to whom is given by a scale-free small-world network. The latter may be a more realistic approximation of human society, albeit still a strong abstraction from actual social interaction patterns.

Both experiments were executed as follows. We first generated a random influence matrix P, conforming to either basic interaction structure. We then ran each of the four strategies we described above on each P and recorded the population opinion at each of 100 rounds of updating.

Interaction networks. In contrast to the influence matrix P, which we can think of as the adjacency matrix of a directed and weighted graph, we model the basic interaction structure of a population, i.e., the qualitative structure that underlies P, as an undirected graph $G = \langle N, E\rangle$ where $N = \{1, \ldots, n\}$ is the set of nodes, representing the agents, and $E \subseteq N \times N$ is a reflexive and symmetric relation on N.[21] If $\langle i, j\rangle \in E$, then, intuitively speaking, i and j know each other, and either agent could in principle influence the opinion of the other. For each agent i, we consider $N(i) = \{j \in N \mid \langle i, j\rangle \in E\}$ the set of i's *neighbors*. The number of i's neighbors is called agent i's *degree* $d_i = |N(i)|$. For convenience, we will restrict attention to *connected* networks, i.e., networks all of whose nodes are connected by some sequences of transitions along E. Notice that this also rules out agents without neighbors.

For a homogeneous society, as modelled in our first experiment, we assumed that the interaction structure is given by a totally connected graph. For heterogeneous societies, we considered so-called *scale-free small-world networks* (Barabási and Albert, 1999; Albert and Barabási, 2002). These networks are characterized by three key properties which suggest them as somewhat realistic models of human societies (c.f. Jackson, 2008):

[21] Normally social network theory takes E to be an irreflexive relation, but here we want to include all self-connections so that it is possible for all agents to be influenced by their own opinion as well.

(1.) *scale-free*: at least some part of the distribution of degrees has a power law character (i.e., there are very few agents with many connections, and many with only a few);

(2.) *small-world*:

  (a.) *short characteristic-path length*: it takes relatively few steps to connect any two nodes of the network (more precisely, the number of steps necessary increases no more than logarithmically as the size of the network increases);

  (b.) *high clustering coefficient*: if j and k are neighbors of i, then its likely that j and k also interact with one another.

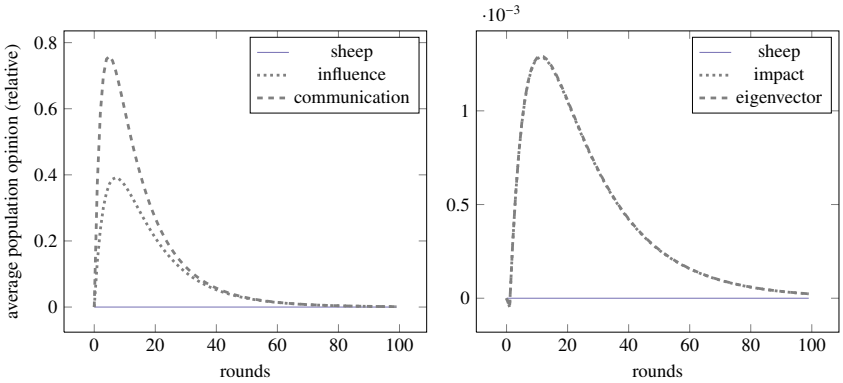
We generated random scale-free small-world networks using the algorithm of Holme and Kim (2002) with parameters randomly sampled from ranges suitable to produce networks with the above mentioned properties. (We also added all self-edges to these graphs; see Footnote 21.)

For both experiments, we generated graphs of the appropriate kind for population sizes randomly chosen between 100 and 1000. We then sampled a number of wolves averaging around 10% of the total number of agents (with a minium of 5) and randomly placed the wolves on the network. Subsequently we sampled a suitable random influence matrix P that respected the basic interaction structure, in such a way that $P_{ij} > 0$ only if $\langle i, j \rangle \in E$. In particular, for each sheep i we independently sampled a random probability distribution (using the r-Simplex algorithm) of size d_i and assigned the sampled probability values as the influence that each $j \in N(i)$ has over i. As mentioned above, we assumed that wolves are unshakably stubborn ($P_{ii} = 1$).

Results. For the most part, our experiments vindicated our expectations about the four different strategies that we tested. But there were also some interesting surprises.

The temporal development of average relative opinions under the relevant strategies is plotted in Figure 5 for homogeneous societies and in Figure 6 for heterogeneous societies. Our general expectation that strategies `influence` and `communication` are good choices for fast success after just a few rounds of play is vindicated for both types of societies. On the other hand, our expectation that targeting influential players with strategies `impact` and `eigenvector` will be successful especially in the long run did turn out to be correct, but only for the heterogeneous society, not for the homogeneous one. As this is hard to see from Figures 5 and 6, Figure 7 zooms in on the distribution of relative opinion means at the 100$^{\text{th}}$ round of play.

At round 100 relative means are very close together because population opinion is close to wolf opinion already for all strategies. But even though the relative opinions at the 100$^{\text{th}}$ round are small, there are nonetheless significant differences. For homogeneous societies we find that *all* means of relative opinion at round 100 are significantly different ($p < .05$) under a paired Wilcoxon test. Crucially, the difference between `influence` and `impact` is highly significant ($V = 5050$, $p < .005$). For the heterogeneous society, the difference between `influence` and `impact` is also significant ($V = 3285$, $p < 0.01$). Only the means of `communication` and `influence` turn out not significantly different here.



(a) Strategies targeting influenceable sheep

(b) Strategies targeting influential sheep

Figure 5: Development of average population opinion in homogeneous societies (averaged over 100 trials). The graph in Figure 5a shows the results for strategies targeting influenceable sheep, while one in Figure 5b shows strategies targeting influential sheep. Although curves are similarly shaped, notice that the *y*-axes are scaled differently. Strategies `influence` and `communication` are *much* better than `impact` and `eigenvector` in the short run.

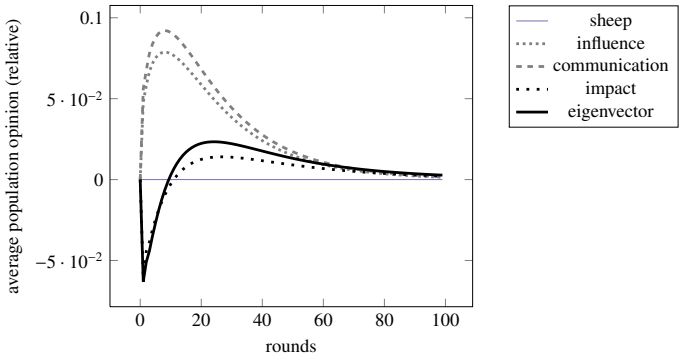


Figure 6: Development of average relative opinion in heterogeneous societies (averaged over 100 trials).

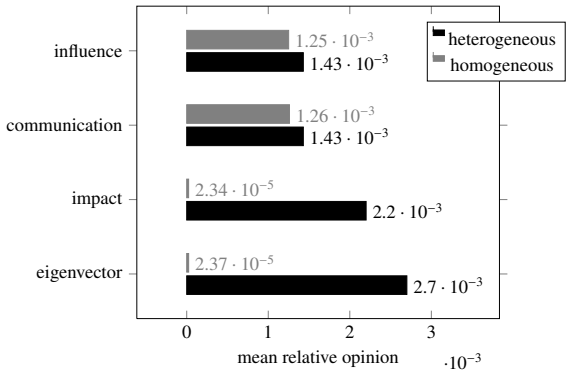


Figure 7: Means of relative opinion at round 100 for heterogeneous and homogeneous societies. Strategies `impact` and `eigenvector` are efficient in the long run in heterogeneous societies with a pronounced contrast between more and less influential agents.

Indeed, contrary to expectation, in homogeneous societies strategies preferentially targeting influenceable sheep were more successful on average for *every* $0 < k \leq 100$ than strategies preferentially targeting influential sheep. In other words, the type of basic interaction structure has a strong effect on the success of a given (type of) manipulation strategy. Although we had expected such an effect, we had not expected it to be that pronounced. Still, there is a plausible *post hoc* explanation for this observation. Since in homogeneous societies (almost) every wolf can influences (almost) every sheep, wolves playing strategies `impact` and `eigenvector` invest effort (almost) exactly alike. But that means that most of the joint effort invested in influencing the same targets is averaged out, because everybody heavily invests in these targets. In other words, especially for homogeneous societies playing a coalition strategy where manipulators do not get into each other's way are important for success. If this explanation is correct, then a very interesting practical advice for social influencing is ready at hand: given the ever more connected society that we live in, with steadily growing global connectedness through telecommunication and social media, it becomes more and more important for the sake of promoting one's opinion within the whole of society to team-up and join a coalition with like-minded players.

3 Conclusions, related work & outlook

This paper investigated strategies of manipulation, both from a pragmatic and from a social point of view. We surveyed key ideas from formal choice theory and psychology to highlight what is important when a single individual wants to manipulate the choice and opinion of a single decision maker. We also offered a novel model of strategically influencing social opinion dynamics. Important for both pragmatic and social aspects of manipulation were heuristics, albeit it in a slightly different sense here and there: in order to be a successful one-to-one manipulator, it is important to know the

heuristics and biases of the agents one wishes to influence; in order to be a successful one-to-many manipulator, it may be important to use heuristics oneself. In both cases, successful manipulation hinges on exploiting weaknesses in the cognitive make-up of the to-be-influenced individuals or, more abstractly, within the pattern of social information flow. To promote an opinion in a society on a short time scale, one would preferably focus on influenceable individuals; for long-term effects, the focus should be on influential targets.

The sensitivity to framing discussed in Section 1.3 is akin to traditional fallacies of argumentation (Hamblin, 1970): hearers will be influenced not by the content of the message, but by the way it is communicated. In contrast to the standard literature on fallacies, our focus in this paper was not on their invalidity, but on the question what it is about us that makes these fallacious inferences so common. We have argued in this paper that this is to a large extent due to agents' limited abilities. Hahn and Oaksford (2007) as well as Mercier and Sperber (2011) argue instead that although we are able to reason very well (to persuade ourselves or others of a particular position), our reasoning is imperfect especially when stakes and targets are low. In that case, it seems, the reasoning is not really fallacious, but rather is one that normally will do. In fact, this is not so different from the reasons we mentioned why agents might sometimes be manipulated: they reason in a way that normally works well. Non-monotonic logics (including probabilistic logic) are natural tools to account for these types of inferences (Gabbay and Woods, 2008). Indeed, non-monotonic reasoning seems a natural resource-compensation strategy, next to approximation and the division of the world into natural kinds. Although this reasoning is not always accurate, it makes perfect sense when balancing the cost of calculation versus the potential benefit of the result. In contrast to these reasons, however, we pointed out that even in argumentative situations, where hearers reason well, speakers can still manipulate by obfuscating states of the world that are to the speaker's disadvantage, and of which the hearer is unaware; or by making the hearer selectively aware of states of the world of which he was previously unaware, i.e., only of those which are to the sender's advantage.

Many important features of strategic manipulation have not been addressed and must be left for future work. Most strikingly, the model of social influencing given in the second part of the paper is heavily simplistic in a number of ways. We have not at all addressed the case where several manipulators with different motives compete for influence over the population opinion. In that case, we would really consider a full game problem where what is a good manipulation strategy also depends on what competing manipulators do. We have also assumed that the pragmatic one-to-one aspect of opinion manipulation does not play a role when it comes to the social problem of opinion manipulation. Of course, there are obvious interactions: the social structure of the population will likely also affect *which* information to present to *whom* and *how* to present information to *this* or *that* guy. To the best of our knowledge, this is largely uncharted terrain for formal theories of strategic manipulation. Adding the broader social perspective to models of social influencing in formal choice models seems a very promising field for future research. A lot of work would need to be done here to tackle the immense complexity of this subject. The model presented in the second half of this paper might be a first step in this direction.

Relation to other work presented in this volume. Gabriel Sandu's work that is presented in this volume appears to be closely related to our own, but there are fundamental conceptual differences. Sandu works in the tradition of Hintikka's game semantics. Game semantics is an approach to formal semantics that seeks to ground the concepts of truth, meaning, and validity in terms of a dialogical view of communication and of winning strategies. While Sandu's approach is primarily about *semantics*, and the grounding of truth, we focus on (perhaps non-Gricean) *pragmatics*, on how agents can be influenced by means of communication. Whereas in Sandu's tradition the dialogues can be very long, but the meaning of the dialogue moves is always clear, the dialogues we studied were typically short, but vague, in that is might be hard to determine what was meant by a dialogue move. The roles of the participants of the dialogues is very different as well. In Sandu's semantic games, there are only two participants with fixed goals: either to verify or falsify a given formula. It is common knowledge between the participants that their goals are diametrically opposed. Although when we talk about one-to-one communication, we also limit ourselves to communication games with only two participants involved, the emphasis of the later part of our paper is on influencing whole groups of agents. Equally important, the goals of our agents are not necessarily diametrically opposed but can show various degrees of alignment.

There are other interesting connections between this paper and others presented in this volume. Eric Pacuit explores a number of ways to reason strategically about a game situation. This links directly to the first part of this paper, where we argued that knowing a decision maker's cognitive make-up opens possibilities of exploitation by a malignant communicator. Taking Pacuit's and our perspective together, a very interesting open question arises, namely which reasoning strategies (in the sense of Pacuit) are more or less vulnerable to malign influence by a strategic self-interested communicator.

Andrés Perea acknowledges, like we do, the intuitive necessity to look for more cognitively realistic ways of solving games. But Perea's conceptual approach is a slightly different one from ours. While we are interested in integrating concrete psychological aspects of reasoning, Perea's contribution remains more conceptual in that it shows, roughly speaking, that a normatively compelling solution of a game problem can be reached at less cognitive effort than frequently assumed.

Finally, Jan van Eijck and Floor Sietsma discuss many different aspects of how the strategic reasoning of individuals can impact the well-fare of society as a whole. Our second model can well be seen as one special case of this more general perspective. Adding their approach to social strategizing to ours raises another interesting open issues that we have not explicitly addressed. The model of strategic influencing that we presented here did not actually specify whether the influence exerted by the strategic manipulators was beneficial or detrimental to society as a whole. Another related issue, brought up by Gabriel Sandu (p.c.), relates to the question of truth. What if the social opinion dynamics are not only affected by what others believe, but also by what is good and/or true?

In summary, unifying psychological and social aspects of manipulative strategizing in a unified framework is a giant's task that has close bearing on many central concerns raised in other papers of this volume. We see a lot of potential for future research in this area, especially where it brings formal modelling and empirical research closer

together.

Acknowledgements. This paper profited from many insightful comments of the participants of the Workshop *Modeling Strategic Reasoning*. We are particularly thankful for the stimulating input from Rineke Verbrugge, Johan van Benthem, Anton Benz, Gabriel Sandu and Jakub Szymanik.

References

Acemoglu, Daron and Asuman Ozdaglar (2011). "Opinion Dynamics and Learning in Social Networks". In: *Dynamic Games and Applications* 1.1, pp. 3–49.

Albert, Réka and Alber-László Barabási (2002). "Statistical Mechanics of Complex Networks". In: *Reviews of Modern Physics* 74.1, pp. 47–97.

Axelrod, Robert (1997). "The Dissemination of Culture: A Model with Local Convergence and Global Polarization". In: *Journal of Conflict Resolution* 41.2, pp. 203–226.

Barabási, Alber-László and Réka Albert (1999). "Emergence of Scaling in Random Networks". In: *Science* 286.5439, pp. 509–512.

Benz, Anton (2006). "Utility and Relevance of Answers". In: *Game Theory and Pragmatics*. Ed. by Anton Benz et al. Hampshire: Palgrave, pp. 195–219.

— (2007). "On Relevance Scale Approaches". In: *Proceedings of Sinn und Bedeutung 11*. Ed. by Estela Puig-Waldmüller, pp. 91–105.

Benz, Anton and Robert van Rooij (2007). "Optimal Assertions and what they Implicate". In: *Topoi* 26, pp. 63–78.

Camerer, Colin F. (2003). *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton University Press.

Camerer, Colin F. et al. (2004). "A Cognitive Hierarchy Model of Games". In: *The Quarterly Journal of Economics* 119.3, pp. 861–898.

Castellano, Claudio et al. (2009). "Statistical Physics of Social Dynamics". In: *Reviews of Modern Physics* 81, pp. 591–646.

Crawford, Vincent P. (2003). "Lying for Strategic Advantage: Rational and Boundedly Rational Misrepresentation of Intentions". In: *American Economic Review* 93.1, pp. 133–149.

— (2007). "Let's Talk It Over: Coordination Via Preplay Communication With Level-k Thinking". Unpublished manuscript.

DeGroot, Morris H. (1974). "Reaching a Consensus". In: *Journal of the American Statistical Association* 69.345, pp. 118–121.

Farrell, Joseph (1988). "Communication, Coordination and Nash Equilibrium". In: *Economic Letters* 27.3, pp. 209–214.

— (1993). "Meaning and Credibility in Cheap-Talk Games". In: *Games and Economic Behavior* 5, pp. 514–531.

Farrell, Joseph and Matthew Rabin (1996). "Cheap Talk". In: *The Journal of Economic Perspectives* 10.3, pp. 103–118.

Feinberg, Yossi (2008). “Meaningful Talk”. In: *New Perspectives on Games and Interaction*. Ed. by Krzysztof R. Apt and Robert van Rooij. Amsterdam: Amsterdam University Press, pp. 105–119.
— (2011a). “Games with Unawareness”. Unpublished manuscript, Stanford University.
— (2011b). “Strategic Communication”. In: *Proceedings of the 13th Conference on Theoretical Aspects of Rationality and Knowledge*. Ed. by Krzysztof R. Apt. New York: ACM, pp. 1–11.
Franke, Michael (2009). “Signal to Act: Game Theory in Pragmatics”. PhD thesis. Universiteit van Amsterdam.
— (2010). “Semantic Meaning and Pragmatic Inference in Non-cooperative Conversation”. In: *Interfaces: Explorations in Logic, Language and Computation*. Ed. by Thomas Icard and Reinhard Muskens. Lecture Notes in Artificial Intelligence. Berlin, Heidelberg: Springer-Verlag, pp. 13–24.
— (2011). “Quantity Implicatures, Exhaustive Interpretation, and Rational Conversation”. In: *Semantics & Pragmatics* 4.1, pp. 1–82.
— (forthcoming). “Pragmatic Reasoning about Unawareness”. In: *Erkenntnis*.
Franke, Michael et al. (2012). “Relevance in Cooperation and Conflict”. In: *Journal of Logic and Computation* 22.1, pp. 23–54.
Gabbay, Dov and John Woods (2008). “Resource-origins of nonmonotonicity”. In: *Studia Logica* 88, pp. 85–112.
Gilboa, Itzhak and David Schmeidler (2001). *A Theory of Case-Based Decisions*. Cambridge University Press.
Grice, Paul Herbert (1975). “Logic and Conversation”. In: *Syntax and Semantics, Vol. 3, Speech Acts*. Ed. by Peter Cole and Jerry L. Morgan. Academic Press, pp. 41–58.
Hahn, Ulrike and Michael Oaksford (2007). “The rationality of informal argumentation: A Bayesian approach to reasoning fallacies”. In: *Psychological Review* 114.3, pp. 704–732.
Halpern, Joseph Y. and Leandro Chaves Rêgo (2006). “Extensive Games with Possibly Unaware Players”. In: *Proceedings of the Fifth International Joint Conference on Autonomous Agents and Multiagent Systems*, pp. 744–751.
Hamblin, Charles L. (1970). *Fallacies*. London: Methuen.
Hegselmann, Rainer and Ulrich Krause (2002). “Opinion Dynamics and Bounded Confidence: Models, Analysis, and Simulation”. In: *Journal of Artificial Societies and Social Simulation* 5.3.
Heifetz, Aviad et al. (2012). “Dynamic Unawareness and Rationalizable Behavior”. Unpublished manuscript.
Holme, Petter and Beom Jun Kim (2002). “Growing scale-free networks with tunable clustering”. In: *Physical Review E* 65.2, pp. 026107–1–026107–4.
Jackson, Matthew O. (2008). *Social and Economic Networks*. Princeton University Press.
Jäger, Gerhard (2011). “Game-Theoretical Pragmatics”. In: *Handbook of Logic and Language*. Ed. by Johan van Benthem and Alice ter Meulen. Amsterdam: Elsevier, pp. 467–491.

Jäger, Gerhard and Christian Ebert (2009). "Pragmatic Rationalizability". In: *Proceedings of Sinn und Bedeutung 13*. Ed. by Arndt Riester and Torgrim Solstad, pp. 1–15.

Kahane, Howard and Nancy Cavender (1980). *Logic and Contemporary Rhetoric*. Belmont: Wadsworth Publishing.

Kahnemann, Daniel and Amos Tversky (1973). "On the Psychology of Prediction". In: *Psychological Review* 80, pp. 237–251.

Lehrer, Keith (1975). "Social consensus and rational agnoiology". In: *Synthese* 31.1, pp. 141–160.

Levinson, Stephen C. (1983). *Pragmatics*. Cambridge, UK: Cambridge University Press.

Lewis, David (1969). *Convention. A Philosophical Study*. Cambridge, MA: Harvard University Press.

Luce, Duncan R. (1959). *Individual Choice Behavior: A Theoretical Analysis*. New York: Wiley.

Matthews, Steven A. et al. (1991). "Refining Cheap Talk Equilibria". In: *Journal of Economic Theory* 55, pp. 247–273.

May, Kenneth O. (1945). "Intransitivity, Utility and the Aggregation of Preference Patterns". In: *Econometrica* 22.1, pp. 1–13.

Mercier, Hugo and Dan Sperber (2011). "Why do humans reason? Arguments from an argumentative theory". In: *Behavioral and Brain Sciences* 34, pp. 57–111.

Milgrom, Paul and John Roberts (1986). "Relying on the Information of Interested Parties". In: *RAND Journal of Economics* 17.1, pp. 18–32.

Myerson, Roger B. (1989). "Credible Negotiation Statements and Coherent Plans". In: *Journal of Economic Theory* 48.1, pp. 264–303.

O'Keefe, Daniel J. and Jakob D. Jensen (2007). "The relative persuasiveness of gain-framed and loss-framed messages for encouraging disease prevention behaviors". In: *Journal of Health and Communication* 12, pp. 623–644.

Ozbay, Erkut Y. (2007). "Unawareness and Strategic Announcements in Games with Uncertainty". In: *Proceedings of TARK XI*. Ed. by Dov Samet, pp. 231–238.

Parikh, Prashant (1991). "Communication and Strategic Inference". In: *Linguistics and Philosophy* 473–514.14, p. 3.

— (2001). *The Use of Language*. Stanford University: CSLI Publications.

— (2010). *Language and Equilibrium*. MIT Press.

Rabin, Matthew (1990). "Communication between Rational Agents". In: *Journal of Economic Theory* 51, pp. 144–170.

Rogers, Brian W. et al. (2009). "Heterogeneous Quantal Response Equilibrium and Cognitive Hierarchies". In: *Journal of Economic Theory* 144.4, pp. 1440–1467.

van Rooij, Robert and Michael Franke (2012). "Promises and Threats with Conditionals and Disjunctions". In: *Discourse and Grammar: From Sentence Types to Lexical Categories*. Ed. by Günther Grewendorf and Thomas Ede Zimmermann. Berlin: de Gruyter Mouton, pp. 69–88.

van Rooy, Robert (2003). "Quality and Quantity of Information Exchange". In: *Journal of Logic, Language and Computation* 12, pp. 423–451.

Shin, Hyun Song (1994). "The Burden of Proof in a Game of Persuasion". In: *Journal of Economic Theory* 64.1, pp. 253–264.

Simon, Herbert A. (1959). “Theories of decision-making in economics and behavioral science”. In: *American Economic Review*.

Stalnaker, Robert (2006). “Saying and Meaning, Cheap Talk and Credibility”. In: *Game Theory and Pragmatics*. Ed. by Anton Benz et al. Hampshire: Palgrave MacMillan, pp. 83–100.

Tversky, Amos and Daniel Kahnemann (1974). “Judgement under Uncertainty: Heuristics and Biases”. In: *Science* 185, pp. 1124–1131.

— (1981). “The Framing of Decisions and the Psychology of Choice”. In: *Science* 211.4481, pp. 453–458.

Zapater, Inigo (1997). “Credible Proposals in Communication Games”. In: *Journal of Economic Theory* 72, pp. 173–197.