Andreas Engert[*]

Norms, Rationality, and Communication: A Reputation Theory of Social Norms

Does the discovery of "law and social norms" necessitate breaking with the rational choice paradigm? In this paper, I argue for an answer in the negative. To this end, I propose a reputation theory of social norms, which differs from other proposals in two principal respects: First, it explains norms without any assumption of behavioral constraints (like habit or conscience) and normative motivations (like altruism or aspiration to esteem). Second, it does even without any assumption regarding model-exogenous, private information that most other reputation and signaling explanations use (such as the discount rate in Eric Posner's signaling model).

Instead, reputation theory analyzes norms as mere social constructs: In strategic situations, rationality fails to provide clear guidance on how to play. Because individuals must nonetheless make decisions, they follow norms. Yet norms are susceptible to strategic manipulation; they can be destabilized by promulgating different norms. Two factors make norms stable in spite of that threat: network effects and preference compatibility. These two factors favor cooperative reputation norms, i.e., norms that foster exchanges among few individuals. More excitingly, network effects and preference compatibility also support norms that overcome collective action problems. Thus, the reputation theory of social norms is an additional way of resolving one of the anomalies of rational choice analysis – the fact that collective action exists.

[*] Institute for International Law, Munich University, Veterinärstr. 5, 81543 München, Germany; phone +49 171 4767967; e-mail andreas@engert.info. – I wish to thank Cass Sunstein (who supervised the writing of an earlier version as independent research project at the University of Chicago Law School). I have benefited greatly from comments by Gary Becker, Eric Posner, Jacob Hanisch, and Julia Pohlmeier on earlier drafts. Last not least, I am grateful to Jörg Wulftange and Steffen Wesche for our ongoing exchange of ideas on the subject. No need to say, all mistakes are mine.

Norms, Rationality, and Communication: A Reputation Theory of Social Norms

Law and economics has given birth to another school of thought, *law and (social) norms*.[1] Family relations tend to be conflict-laden. While parents aspire to leave an imprint perpetuating their personality and values, children often grow mature only by opposing to those views. Law and norms has been pretty much mummy's darling before going through a phase of defiance.[2] Still, many believe that law and norms has not yet gained a lot of maturity.[3] Seemingly, it has not developed much of a personality beyond the interest in its subject – norms.[4] In this paper, I consider one, perhaps the principal, question to define law and norm's character. Little surprisingly,

---

[1] See generally the commenty by Eric A. Posner, The Signaling Model of Social Norms: Further Thoughts, 36 U. Rich. L. Rev. 465 (2002); Robert C. Ellickson, Law and Economics Discovers Social Norms, 27 J. L. S. 537 (1998), and Richard A. Posner, Social Norms, Social Meaning, and Economic Analysis of Law: A Comment, 27 J. L. S. 553, 564-565 (1998).

[2] The latter might be associated with the activist "New Chicago School", Lawrence Lessig, The New Chicago School, 27 J. L. S. 661 (1998); Cass Sunstein, Social Norms and Big Government, 15 Quinnipiac L. Rev. 147 (1995), Cass R. Sunstein, Social Norms and Social Roles, 96 Col. L. Rev. 903 (1996), particularly at 907 ("norm management") and 947 *et seq.*, and Dan M. Kahan, Social Influence, Social Meaning, and Deterrence, 83 Va. L. R. 349 (1997). A recent, well-weighed assessment can be found in Robert C. Ellickson, The Market for Social Norms, 3 Am. L. & Econ. Rev. 1, 30-42 (2001).

[3] Recently, skeptical voices made themselves heard. See Robert E. Scott, The Limits Of Behavioral Theories Of Law And Social Norms, 86 Va. L. R. 1603 (2000). For its limited use in the theory of the firm see Oliver Hart, Norms and the Theory of the Firm, 149 Penn. L. Rev. 1701 (2001).

[4] See Eric A. Posner, The Signaling Model of Social Norms: Further Thoughts, 36 U. Rich. L. Rev. 465, 465 (2002).

it is at the same time a core component of law and economics: I am speaking of rational choice.[5]

Law and norms can accept or waive the heritage of rational choice. In this paper, I undertake to defend rational choice as a solid foundation for positive law and norms theory. To this end, I propose a *reputation theory of social norms*. The reputation theory of social norms avoids assumptions about behavioral traits, preferences, or internal constraints. That is, it does without many assumptions that have been used to account for social norms, such as the notion of internalization or normative motivations.[6] However, the reputation theory put forward here is even

---

[5] For the critique of rational choice in the domain of law and economics itself see Christine Jolls *et al.*, A Behavioral Approach to Law and Economics, 50 Stan. L. Rev. 1471 (1998) and the other articles compiled in Cass R. Sunstein (ed), Behavioral Law and Economics (2000).

[6] On internalization, see only Robert D. Cooter, Do Good Laws Make Good Citizens? : An Economic Analysis of Internalized Norms, 86 Va. L. R. 1577 (2000); Robert D. Cooter, Expressive Law and Economics, 27 J. L. S. 585 (1998); Gary S. Becker, Norms and the Formation of Preferences, in Accounting for Tastes 225 (2 ed 1998); Robert H. Frank, If Homo Economicus Could Choose His Own Utility Function, Would He Want One With a Conscience?, 77 Am. Econ. Rev. 593 (1987). On the preference side, it has been argued that people tend to have a preference for fairness and reciprocity (Gary S. Becker, Norms and the Formation of Preferences, in Accounting for Tastes 225 (2 ed 1998); Lawrence Lessig, The Regulation of Social Meaning, 62 Chicago L. Rev. 943, 1003 (1995)) or for other normative values. It is in a similar flavor when scholars argue that norms are followed not for instrumental reasons but unconditionally for their own value (Jon Elster, Social Norms and Economic Theory, 3 J. Ec. Persp. 99 (1989); Joseph Heath, The Structure of Normative Control, 17 L. & Philos. 419 (1998)), out of altruistic preferences (Lynn A. Stout, Other-Regarding Preferences and Social Norms, Georgetown Law and Economics Research Paper No 265902, (Washington, 2001)), out of emotions (Ernst Fehr & Simon Gächter, Altruistic punishment in humans, 415 Nature 137 (2002)), for conformity (Douglas B. Bernheim, A Theory of Conformity, 102 J. Pol. Ec. 841 (1994)), or for being esteemed (Richard H. McAdams, The Origin, Development, and Regulation of Norms, 96 Mich. L. Rev. 338, 355 *et seq.* (1997)). On the constraint side, the idea usually is that people incur some kind of

more radical than that.[7] While other approaches, notably E. Posner's signaling theory, have also relied on reputation, they have insisted that an individual's reputation conveys a piece of information on her character, preferences, or other intrinsic property.[8] By contrast, the reputation theory of norms advocated here goes without even such modest assumptions, thus complementing other rational choice accounts: Reputation is conceived of as a mere social construct. The underlying claim is that there can be a promising rational choice theory of social norms, and that reputation theory is at least part of such a comprehensive approach.

The paper contains five sections. The first section introduces the notion of rational indeterminacy. The concept of norms is defined as a response to rational indeterminacy. The second section raises the problem of norm stability as the key challenge to norms. A solution is developed at the general level using the notions of norm network effects and preference compatibility. In the third section, the solution concept is applied at a more concrete level to explain the stability of reputation norms. Based on a better understanding of norm stability, it is shown that reputation is strong enough to induce even compliance with norms in favor of collective goods. Also, reputation theory is compared to the norm theories by McAdams and by E. Posner. Up to this point, the paper is of a very theoretical kind. Therefore, the fourth

---

internal cost or punishment when violating the norm (Robert D. Cooter, Expressive Law and Economics, 27 J. L. S. 585, 597 *et seq.* (1998); Richard A. Posner & Eric Rasmusen, Creating and Enforcing Norms, with Special Reference to Sanctions, 19 Intl. L. & Econ. Rev. 369 (1999)).

[7] Cf. the critique of E. Posner's reductionism in Richard H. McAdams, Signaling Discount Rates: Law, Norms, and Economic Methodology, 110 Y. L. J. 625, 681-6 (2001).

section offers a few examples. The fifth section concludes with a short outlook, particularly for normative analysis.

As even this short overview reveals, the paper dwells quite a while on issues that seem to have little to do with reputation. Those arguments, relating to norm stability and communication, are important parts of the theory. Nonetheless, it is warranted to speak of a *reputation* theory of norms. All those arguments help make the case that reputation is a richer and more powerful concept than is usually believed.

I. Norms and Rational Indeterminacy

The central problem of a norms theory under methodological individualism and rational choice is: How can cooperation and collective action evolve in view of individual utility maximization? I will argue that, paradoxically, the solution to this problem lies in the fact that rational individuals (or, as I will say: players) have to cope with an even more fundamental challenge, namely the problem of *rational indeterminacy*. Basically, rational indeterminacy is the problem that when all actors are rational it becomes hard to tell how to act rationally.

Economics has taught us a lot about cooperative relationships. Under *cooperation*, I understand all kinds of exchanges or trades between very few people – the kind of trades that conventionally are deemed less problematic than *collective*

---

[8] In E. Posner's model, this is the individual discount rate, see III.4. *infra* at 43

action involving many players. The most famous cooperation game is the Prisoner's Dilemma (Table 1).

	a	**b**
A	2 / 2	0 / 3
B	3 / 0	1 / 1

TABLE 1 – PRISONER'S DILEMMA

As is well known, there is a solution to the dilema: A cooperative outcome can be attained if the game is played not just once, but is repeated infinitely.[9] With the prospect of infinite repetition, players can agree on Tit for Tat or a similar strategy. They can induce each other to play the cooperative move by promising to play cooperatively in future rounds, which promise in turn is kept because of future rounds, and so on *ad infinitum*.

[9] An additional condition is that the discount factor must be sufficiently close to 1, i.e. the players must value the future sufficiently high. For a full statement of the "Folk Theorem" see Drew Fudenberg & Jean Tirole, Game Theory 150 *et seq.* (1996). Duncan Luce & Howard Raiffa, Games and Decisions : Introduction and Critical Survey 94 *et seq.* (1957) provide an illustrative discussion of the finitely repeated version and its quandaries.

This mechanism seems so obvious that not many questions have been asked. But strategies like Tit for Tat are more problematic than is commonly believed. Consider the standard setting of the repeated Prisoners' Dilemma. As we have seen (and is intuitively clear), a cooperative outcome turns on a threat not to cooperate in future rounds if the partner flunks in the present round. Threatening implies *conditional play*, that is, players must condition their decisions on what has happened in the past. Natural as this requirement is, it is a problem for rational players. Rational players cannot commit to one long-term strategy, once and forever. Instead, they are free in every round to disregard their long-term strategy.[10] In particular, whether row-player (in Table 1) should play her cooperative move in round 1 hinges on whether column player is going to play conditionally in round 2. What should row-player believe? Suppose she believes that column-player will play conditionally. Is there any incentive for column-player to play in accordance with this belief? Without further assumptions, there is no such incentive: For in round 2, column-player can no longer influence the outcome in round 1. All that matters to her is round 2 and the following rounds. In principle, it would be perfectly rational for column-player to be cheated in every round and, nevertheless, play on cooperatively because row-player promises to

[10] This problem has seldom been analyzed explicitly in this context. It is noticed, *inter alia*, in David M. Kreps, Game Theory and Economic Modelling 71 (1990) and Jordan Howard Sobel, Utility Maximizers in Iterated Prisoners' Dilemmas, in Campbell & Sowden, ed, Paradoxes of Rationality and Cooperation 306, 311 *et seq.* (1985). Binmore argues that if a player observes her opponent deviating in spite of her credible threat she might lose trust in her opponent's rationality and thus refrain from cooperating. See Ken Binmore, 2 Game Theory and the Social Contract : Just Playing 356 (1998).

play cooperatively from now on.[11] The Prisoner's Dilemma has turned into a Punisher's Dilemma.[12]

The Punisher's Dilemma arises even though the players are assumed to be utility maximizers. I should rather say: It arises *because* the players are utility maximizers (and nothing more).[13] To some rational indeterminacy may appear too counterintuitive – an intriguing puzzle perhaps, but not one that helps explain social reality. Particularly, it seems as if rational indeterminacy rested on a very strong, unrealistic concept of rationality. However, consider a milder interpretation: Assume that individuals follow a behavioral program but that this program is of such an enormous complexity that it is by no means possible to predict it. If the behavioral program is to maximize the individual's utility, then rational indeterminacy will again pop up in designing the program. Somewhat simpler put, even if you do not buy rational indeterminacy as a problem of whether conditional play can exist *at all*, it is a much more difficult to escape the problem of predicting the concrete conditions that

---

[11] Intuitively, when cooperation has collapsed for some time players will make attempts to re-start cooperation. The game theoretic literature on "renegotiation proofness" predicts that they will return to cooperation after a cheater has been punished to an extent that would have deterred her *ex ante* (cf. Drew Fudenberg & Jean Tirole, Game Theory 179-82 (1996); Roger B. Myerson, Game Theory : Analysis of Conflict 408-12 (3 ed 1997)). The very point of this literature is that cheating must not be a profitable strategy in the first place. This, however, is exactly what the above argument casts doubt on: The very question is if rational maximizers of utility can attain cooperation *at all*.

[12] Usually, the term is reserved for collective good problems, that is for the problem of how players can be motivated to incur the cost of punishment when others fail to contribute to the collective good. Therefore, Buchanan distinguishes the above problem by calling it the "time dimension" of the punishment dilemma, see James Buchanan, The Limits of Liberty 134 (1975).

lend themselves to conditional play. Suppose player *A* has cheated on player *B*. Which inferences can be drawn? Sometimes, authors postulate that what happened in the past is likely to happen again in the future.[14] One pretty natural inference then is that *A* will cheat again on *B* in the same type of transaction. But what is the same type of transaction? Do we expect murder from a shoplifter? Further: Does *C* has to conclude from *A's* conduct towards *B* that *A* is going to cheat on her as well? What if *C* has different business with *A* or belongs to a different group?

The answer to these questions is likely to vary greatly over situations, groups, and cultures. Under the theory presented here, it is social norms that fill the vast space opened up by rational indeterminacy, thus causing such variations. This, of course, is only the description of norms' function. What interests us more is how this function is fulfilled by norms and, accordingly, what norms actually consist of. The answer I offer to this question is simple. In my view, the empty space of indeterminacy is conquered by the softest of all forces, information. More specifically, it is information generated mostly by means of communication.

---

[13] The argument in the text is closely related to the concept of double contingency as set out in Niklas Luhmann, Social Systems 103 *et seq.* (1995).

[14] This claim seems somewhat naive for epistemological reasons that are of quite practical relevance. Roughly, the question is: What exactly does it mean to act "in the same way?" See Robert Sugden, The Role of Inductive Reasoning in the Evolution of Conventions, 17 L. & Philos. 377, 386-7 (1998) and Leif Johansen, Interaction in Economic Theory, 3 Economie Apliquée 229, 245 (1981). The problem is also noticed by E. Posner and compared to legal decision-making. See Eric A. Posner, Efficient Norms, in Newman, ed, 2 The New Palgrave Dictionary of Economics and the Law 19 (1998).

This is quite an astonishing idea. Norms drive people to volunteer for social work, to refuse bribes, and to participate in blood vendettas. It is hard to conceive how mere communication ("cheap talk") can impact on a fact as solid as players' incentive structure.[15] But for a theory that aims at explaining norms without resorting to behavioral constraints (like guilt) or normative preferences (like values) there is not much of an alternative. Thus, I define norms as

(i) an equilibrium strategy combination

(ii) that is common knowledge and

---

[15] Cf. David M. Kreps, Corporate Culture and Economic Theory, in Alt & Shepsle, ed, Perspectives on Positive Political Economy 90, 111 (1990): Reputation is a "decidedly fragile" construction; it works only if it works. – Digging somewhat deeper, the question is if rational players can at all use "any additional information, not incorporated into the utilities of the situation", see David Gauthier, Coordination, 14 Dialogue 193, 210 (1975). While the question is somehow catchy, Gauthier's answer in the negative is unconvincing. It amounts to saying that because there is rational indeterminacy additional information does not matter. Yet this inference would require a further assumption. Specifically, one would have to assume that rationality does not only itself fail to instruct players how to choose but also *precludes* any other prescriptions. While it is true that, trivially, under rational indeterminacy no rational reasons can be given for making or following a *particular* strategy prescription, it does not follow that rational players were bound to disregard *any* (arbitrary, from rationality's point of view) strategy prescription. On the contrary, it is most reasonable for players to take such a strategy prescription into account if it is common knowledge in their community. In this case, the prescription is a mental state of at least two individuals and, from the perspective of each individual, a state of the outside world or a fact. It is not implausible to assume that facts, however fragile they may be, can alter the calculus of rational maximizers of utility. The problem is also discussed by Robert Sugden, Rational Choice: A Survey of Contributions from Economics and Philosophy, 101 Econ. J. 751, 774-8 (1991) and Leif Johansen, Interaction in Economic Theory, 3 Economie Apliquée 229, 244 (1981).

(iii) is commonly expected.[16]

The definition starts out from the game theory's equilibrium concept. A norm must prescribe an equilibrium. If it did not then – by definition – at least one player would deviate. As the other players would anticipate the deviation they would likewise adapt their strategies. The norm would cease to influence behavior.[17] Also, it has become conventional wisdom that for an equilibrium strategy combination to qualify as a norm it must be *common knowledge*, i.e. all players must know the strategy combination, know that all others know it, know that all others know that they know etc. *ad infinitum*.[18] Hence, when a player undertakes to introduce a new

---

[16] The definition bears some resemblance to the one in Cristina Bicchieri, Rationality and Cooperation 232 (1993).

[17] One qualification is necessary. A norm can also function as a starting point for tracing down strategy choices to the choices actually expected. Take, for instance, the norm not to lie, i.e. never to lie. There are plenty of situations where expecting honesty would be rather naive. Still, this does not imply that the norm not to lie is void. Whenever a player incurs a risk of being caught lying and whenever there are sufficiently severe sanctions at hand, other players can with some confidence trust in her being honest. Hence, "never to lie" is a shortcut of the expectations that (i) lies will be sanctioned if possible and (ii) players will be honest where these sanctions suffice. Based on this equilibrium people develop a sense (often implicitly) of when credulity would be misplaced.

[18] Note that common knowledge is not already required by the equilibrium requirement. – The standard reference is David Lewis, Conventions: A Philosophical Study 52 *et seq.* (1969). A strategy description that is not common knowledge can nonetheless be relevant in a situation. For instance, *A* may wish to visit *B* in *B's* office. She knows that *B* usually is in his office at a certain time, so she does not announce her visit. Here, *A's* going to the office at the chosen time is a strategy description that is not common knowledge. However, *A* runs a risk that *B* is on vacation or has an outside appointment. Slightly different, if *B* has been informed of *A's* plan, say, by *A's* spouse without *A's* knowledge, she might cancel her appointment. Still, the intended meeting is not common knowledge. If *A* learns of *B's* appointment (but not of the fact that she has cancelled it), *A* may change her plan. In sum, strategy descriptions that fail to be common knowledge are prone to errors and discoveries. It is for this reason

norm, she needs to create common knowledge of the proposed course of action. She can do so through cheap talk. Alternatively, she can simply play according to the norm that is to be introduced. When a certain move only makes sense under a particular norm then playing the move implies proposing the respective norm. In either case, the proponent of a new course of action has to be anxious about making her proposal common knowledge. Establishing a norm requires making it common knowledge or "communicating" it.[19]

However, common knowledge does not suffice to turn a strategy description into a norm. There can easily be several strategy descriptions that are common knowledge but are, at the same time, inconsistent with each other. A straightforward example is explicit communication of alternative proposals as in a debate or when a question is asked: Should we do *x* or *y*? In such a case, neither *x* nor *y* should be considered a norm. Generally, the definition of norms should not allow, as a regular matter, for the existence of multiple inconsistent norms. For given players in a given situation, it is conceivable that there are more than one *potential* norms but it would

---

that Lewis excluded these "odd cases" from his definition of a convention, see David Lewis, Conventions: A Philosophical Study 59 (1969).

[19] Cf. Marco Colombetti, A Modal Logic of Intentional Communication, 38 Math. Soc. Sciences 171 (1999) for the literal meaning of "communication" as "making something common". – The notion that communication changes matter has been examined by a string of game theoretic literature. The most interesting part of that literature for the present context deals with announcements of future play (rather than an exchange of factual information). See Joseph Farrell, Communication, Coordination and Nash Equilibrium, 27 Econ. Letters 209 (1988) with the strong assumption that players will believe each other unless there is reason for being suspicious. See also Matthew Rabin, A Model of Pre-game Communication, 63 J. Econ. Th. 370 (1994).

be odd to think, again as a regular matter, of more than one *actual* norm. Therefore, an equilibrium strategy combination that is common knowledge is a norm only if, additionally, the relevant players expect it to be actually played.[20] Only equilibrium strategies that are common knowledge and at the same time *commonly expected* qualify as social norms.[21]

The latter stipulation, indispensable as it is in order to limit the scope of the concept, arouses a suspicion of circularity. That a course of action is commonly expected seems a *deus ex machina* to explain any curiosity in the world of norms (not unlike the "explanation" that players just have a taste for complying with norms).[22] To offer a meaningful account, there must be a specific model of when players will in fact expect a particular strategy to be played. The analysis of norm stability in the next section is such a model. It fleshes out the success chances for strategy combinations, both proposed and incumbent ones. Players are skilled, by virtue of their social knowledge, in assessing those chances. Even if they fail initially to anticipate the successful proposals they can adjust their expectations later. The idea thus is that players have a sense of which expectations can become stable, and will adopt only such promising proposals.

---

[20] At least, the players must deem it it possible that the strategy combination will be actually played.

[21] Note that, in analogy to common knowledge, the expectation must also be "common": A norm must be expected to be played, expected to be expected and so on *ad infinitum*.

[22] See Paul G. Mahoney, Norms and Signals: Some Skeptical Observations, 36 U. Rich. L. Rev. 387 (2002) and Robert D. Cooter, Do Good Laws Make Good Citizens? : An Economic Analysis of Internalized Norms, 86 Va. L. R. 1577, 1591 (2000).

II. Stable Norms

Norms are equilibrium strategies that have been made common knowledge. We have seen though that this is not sufficient. As a third condition, I have stipulated that the strategy combination must be commonly expected. I further specified this condition by referring to the (potential) stability of the strategy combination. Now, while there has been intensive research on the equilibrium concept, I am not aware of much analytical work on the persistence of an established equilibrium (i.e., a norm).[23] This is quite astonishing. If norms consist of nothing more than information they must be extremely susceptible to additional information that creates ambiguity and thus tends to restore indeterminacy. I submit it is at this point that the most can be learned about the nature of social norms: Norms must be shaped in a way that assures their survival in spite of abundant incentives to oppose them. I refer to this condition as the *stability condition*. If the stability condition does not obtain, the norm cannot solve even the slightest conflict: As soon as the norm would decide for one party, the other party could produce a different norm – simply by communicating it.[24]

---

[23] Note that my terminology restricts the equilibrium concept. The stability condition could also be included in the concept of equilibrium if the game was modeled more comprehensively. A similar ambiguity – between cheap talk and its costly consequences – is noted in Joseph Farrell, Talk Is Cheap, 85 Am. Econ. Rev. 186, 189-90 (1995).

[24] To be sure, there are norms that do not face opposition, or only very mild opposition as the norm on which side of the road to drive (left-handed players might have a slight preference for driving onthe left hand side). Yet the more interesting cases are certainly those where incentives exist to overthrow the norm. – The problem is akin to Lessig's distinction between contestable and uncontestable kinds of social meaning: A stable norm ascribes a certain meaning to an act, and that the individual actor cannot tamper with that meaning. Cf. Lawrence Lessig, The New Chicago School, 27 J. L. S. 661, 682-3 (1998); similarly Cass R. Sunstein, Social Norms and Social Roles, 96 Col. L. Rev.

Hence, for a norm to persist it does not suffice that it is common knowledge and prescribes an equilibrium. Additionally, the norm must be such that interested players, given existing communication technology and other norms, are not able to establish a different norm in its place. The following subsections introduce two mechanisms that bring about such stability.

1. Stability and Network Effects

The first mechanism protecting the norm consists of a large number of players who would have to opt out of the existing norm and switch to the new norm. To this variable, I refer to as the *network effects* of a norm.[25]

An study of a historical reputation norm among by Avner Greif demonstrates why network effects are decisive. The "Maghribi traders" were a distinct community of Jewish merchants that survived as a group throughout generations and centuries,

---

903, 926-9 (1996). The present theory undertakes to specify what makes a meaning contestable, thus meeting Lessig's demand.

[25] The idea of norm network effects is explicated by Michael Adams, Norms, Standards, Rights, 12 Eur. J. Pol. Econ. 363 (1996). See also Richard A. Posner & Eric Rasmusen, Creating and Enforcing Norms, with Special Reference to Sanctions, 19 Intl. L. & Econ. Rev. 369, 377-80 (1999). See generally on the economics of network effects: William H. Page & John E. Lopatka, Network Externalities, in Bouckaert & De Geest, ed, 1 Encyclopedia of Law and Economics 952 (2000); Nicholas Economides, The Economics of Networks, 14 Intern. J. Ind. Organ. 673 (1996); Paul A. David & Shane Greenstein, The Economics of Compatibility Standards: An Introduction to Recent Research, 1 Econ. Innov. New Techn. 3 (1990). Put somewhat differently, norm stability owes a lot to the collective action problem of switching to a new norm. See Robert D. Cooter, Expressive Law and Economics, 27 J. L. S. 585, 588-97 (1998) and Cass R. Sunstein, Social Norms and Social Roles, 96 Col. L. Rev. 903, 911 (1996).

and in spite of migration. They had developed an especially efficient enforcement mechanism for merchant agents: An agent who cheated became subject to a boycott by all other members of the community. Therefore, a principal who had wished to hire such an agent would have had to provide a higher stream of remuneration to him in order to guarantee his faithfulness. Thus, ironically, a cheater would earn a higher *wage*, that is, a higher remuneration *per transaction*. But because of that higher wage, principals would strictly prefer to hire other, less expensive agents, so that in total the cheater's income would be much lower than that of an honest member of the community. Thus, the reputation norm was self-fulfilling – it was effective because it was effective. Apparently, the prospect of that loss was sufficient to deter most agents from cheating and, as Greif reports, to induce (alleged) cheaters to compensate their principals.

The mechanism used by the Maghribi traders is paradigmatic for the way a (reputation) norm ensures its stability. In Greif's own words:

> It is the uncoordinated response of all the merchants and the interrelations between their expected future behavior and an agent's optimal wage as perceived by an optimal merchant that insures solidarity of incentives. […] Hence, merchants follow the multilateral punishment *despite* the fact that the agent's strategy does not call for cheating any merchant who violated the collective punishment, and *despite* the fact that cheating in the past does not indicate that the agent is a 'lemon.'[26]

[26] Avner Greif, Contract Enforceability and Economic Institutions in Early Trade: The Maghribi Traders' Coalition, 83 Am. Econ. Rev. 525, 534-5 (1993), italics in the original.

A cheating agent in the Maghribi community had little chance to dethrone the reputation norm. The reason is that the norm commanded strong *network effects*: Because all Maghribis relied on the norm, every trader benefited greatly from a spotless reputation and this, in turn, accounted for the price differential between honest agents and cheaters. For a reputation norm like the Maghribis', network effects result from the value of cooperative opportunities that hinge on a good reputation.[27] Just like the network effects inherent to other network goods, the value of reputation increases with the number of players who apply the reputation norm.

The link to stability stems from the cost of uncertainty: Whatever new norm is proposed it is most unlikely that all players will switch, and will switch simultaneously. As a consequence, switching norms entails the risk that there will be two different norms, at least for some time. With two different norms, network effects are weaker: If, by cheating, a player risks her reputation only with part of the population, the reputation mechanism has less leverage. Also, players adhering to the smaller norm network are worst affected, as their norm has the least deterrent effect. Therefore, players are eager to stay with the major norm network.

The only way to push for norm change is to compete against the major norm in terms of focality. But communication cost rises with necessary network size. If a Maghribi trader had intended to violate the norm he would have had to convince at least a great part of community members – who were dispersed all over the

[27] Those benefits stem from participation in a market, as opposed to a small number of

Mediterranean. Hence, strong network effects increase the norm's stability by imposing high communication cost on competing proposals. To amend a norm with strong network effects one has to communicate with a great number of players. What is more, not only must one spread the new norm itself but also must players be convinced that sufficiently many other players are actually going to switch norms.[28]

The magnitude of network effects does not only have implications for a norm's stability but also for the ease with which it is introduced. If a proposed norm would have strong network effects, players are sensitive to whether or not it will become effective. Accordingly, players tend to provide incentives for others to inform them, i.e. to speak out about their own opinion as well as the opinion distribution over the population. However, this interest only exists if the proposal is a serious candidate for becoming the norm. If it is not, players are especially uninterested even if, in substance, they would strongly prefer the norm to be introduced. These self-amplifying (or self-weakening) social dynamics of network effects account for tipping-effects like in the rise and fall of "political correctness".[29]

---

invariable long-term relationships.

[28] This raises an additional question: What convinces players that other players will switch? The situation is pretty clear if they give their explicit assent. If they do not communicate (or if their statement is unknown) there is an ambiguity. As I will argue in the next subsection, in case of an ambiguity players should be expected to play according to the norm proposal they prefer (*infra* at 20). Hence, whether one can count on players to switch norms will, in ambiguous cases, turn on one's beliefs about their preferences.

[29] See Cass R. Sunstein, Social Norms and Social Roles, 96 Col. L. Rev. 903, 912 (1996). For a brief overview of tipping effects in the literature on network effects see only William H. Page & John

2. STABILITY AND PREFERENCE COMPATIBILITY

Network effects make norms stable by virtue of the cost of communication and uncertainty. A stable norm with strong network effects owes its edge to a large number of players who would have to coordinate on a new norm. Yet we have also seen that norm change is not precluded by strong network effects. On the contrary, network effects can promote a new norm if players expect it to succeed. Thus, we need additional criteria for determining which norm proposals are likely to succeed or fail. For instance, there is a strong intuition that purely self-serving aspirations to norm change must fail. A cheater in the Maghribi community most probably cannot succeed with a proposal to exempt only him from losing his reputation even if communication were costless. This example is quite self-evident but a theory of norms should offer an explanation for this intuition.

To that purpose, I extend the analysis to include the players' interest in the substance of a norm. I refer to this interest as *individual preferences*, as opposed to the benefits from adhering to a norm with strong network effects (*network preferences*). Individual preferences result from individual switching costs, which can create a preponderance for the *status quo* (e.g., learning cost).[30] More interestingly,

---

E. Lopatka, Network Externalities, in Bouckaert & De Geest, ed, 1 Encyclopedia of Law and Economics 952, 960-3 (2000).

[30] See Michael Adams, Norms, Standards, Rights, 12 Eur. J. Pol. Econ. 363, 370-2 (1996) and Richard A. Posner & Eric Rasmusen, Creating and Enforcing Norms, with Special Reference to Sanctions, 19 Intl. L. & Econ. Rev. 369, 377-9 (1999). – Another source of switching costs is the fact

individual preferences differ with respect to the content of different norms. To cite my evident example from above, the cheater under the Maghribi norm surely has a strong preference for being pardoned, while all traders have an interest in sustaining the norm's deterrent effect.

If individual preferences are to shape norms we need an idea of how this happens. There are at least three ways in which individual preferences influence the comparative stability of norms.

The first source of influence is that individual players can raise or decrease communication costs for a norm proposal. We have seen that, if a norm has strong network effects, a new proposal must be communicated to as great a number of players as possible. Beside public communication (meetings, mass media etc.), a very important communication channel is conversations between individual players. Because outsiders have difficulties controlling personal communication it gives latitude in what to tell or suppress. Therefore, players can use personal communication to spread only norm proposals that they have individual preferences for.

The second source of influence is related to the first one: Not only do preferred norm proposals spread more easily in personal communication; they also earn (explicit or implicit) approval. When players learn in a number of independent conversations that other players support a proposal they can build confidence that a

---

that a norm change may destroy the reputation of players that stemmed from compliance with the norm.

norm switch is actually going to happen. A social consensus emerges. Conversely, if a norm is disliked players will utter disapproval and announce not to comply with the norm.[31] Because of network effects, a norm proposal that remains controversial is less likely to be relied on for actual behavior.

The third source of influence arises in ambiguous cases: Often, it will not be clear which of several norm proposals is going to succeed. A similar problem occurs when a norm, while not challenged, is ambiguous with regard to a particular situation.[32] Players cannot, in these cases, rely on network effects and settled expectations to make their decision. But in the absence of better criteria they can at least reveal their individual preferences. The way to do that is, simply, to comply with one's preferred norm proposal.

Thus, there are at least three mechanisms – communicating the proposal, expression of consent/dissent, and choice in ambiguous cases – that give more preferred norms a competitive edge over less preferred norms. The process of norm selection is

---

[31] Players can often afford to be honest because they are not bound by their statement. If it turns out that the proposal becomes the norm, they can still adapt to the majority opinion. The only risk they run is a (limited) embarrassment.

[32] E. Posner describes a similar notion by emphasizing the "irreducibly specific" character of "applying" norms to a particular situation, see Eric A. Posner, Efficient Norms, in Newman, ed, 2 The New Palgrave Dictionary of Economics and the Law 19, 20 (1998). The problem is most familiar to every lawyer. See also n. 14 *supra*. In a way, the problems of applying existing norms and introducing new norms are in fact identical: That a norm is ambiguous with regard to a particular situation is little different from saying that there is a lack of a norm deciding the case.

biased towards individually preferred norms. Hence, I will say that, in order to be stable, a norm must be *preference compatible*.[33]

I have argued that stable norms must be preference compatible and must have strong network effects. These two conditions interrelate. Stable norms must at the same time be *shared* and *preferred* by most members of the relevant group. To match both conditions, norms must be an acceptable compromise for most group members. This has an effect on norm design. While trader *A* originally does not care if trader *B* (*C*, *D*, …) is cheated, he is vitally interested that no one cheats on him. Accordingly, *B* does not care for *A*, *C*, *D*, and all other players provided he is not cheated himself. A reputation norm must overcome this indifference. The solution is to bundle the prohibitions to cheat on players *A*, *B*, (*C*, …) to one *general* ban on cheating. Only the general ban is preferred by all traders, can create a large norm network, and thus exploit strong network effects. Now, if a cheater proposed an exception for himself, the ban on cheating would no longer be a general one. Even a single exception would pose a problem. Players will suspect that more exceptions will follow the first, and they may well be right. There is, in other words, a (tacit) norm that no exceptions will

[33] Of course, there is also a variety of reasons why people do not express their preferences in the ways specified in the text. The standard reference is Timur Kuran, Private Truths, Public Lies (1995) but see also Eric A. Posner, Law and Social Norms 186 *et seq.* (2000). It is not quite clear though what this implies for preference compatibility. If speech is restrained by norms, preference compatibility simply returns on a higher level: A norm restraining free speech may very well itself be preference compatible – for the very results it has on "lower order" norms. For the potentially desirable effects of restricting free speech see Dhammika Dharmapala & Richard H. McAdams, Words that Kill: An Economic Perspective on Hate Speech and Hate Crimes, Illinois Law & Econ. Research Paper No 00-34, (Urbana-Champaign, 2001)). Of course, this sketchy analysis by far does not exhaust the problem.

be made.[34] Once loopholes were accepted, the rule would lose its simplicity. Without simplicity, players cannot monitor if others still adhere to the network. If mistrust in the norm's validity grows, the network ultimately collapses. Because players know this consequence, they strongly oppose such norm change.[35]

III. Explaining Norms

Most rational choice theorists feel confident to explain cooperation among few players. What troubles them are norms that go beyond cooperation, namely norms that are costly to players without providing an immediate gain to them.[36] One such puzzle to rational choice theory is that people incur the cost of voting although the chance to make a difference in the election result is vanishing. Another is the fact that people pay more taxes than we would predict considering the expected value of legal sanctions.

---

[34] Of course, real world norms usually have more sophisticated mechanisms to maintain norm confidence. The picture is rough but it lays bare the basic driving force.

[35] This, I think, is the appropriate solution to avoid the fatalism of David M. Kreps, Game Theory and Economic Modelling 71 (1990): Kreps raises the concern that players could find it in their best joint interest to "forgive and forget" in order to preserve opportunities for cooperation with the cheater. Kreps' point is so interesting because it invokes rational indeterminacy and the Punisher's Dilemma (cf. *supra* at 7). The solution is this: With a norm in place, the basic setup of rational indeterminacy has changed. Now, a norm violator has to accomplish an amendment to the norm, which not only takes time and effort but also is likely to fail because other players fear to be left without any norm or with a norm that is less effective. In many cases, these factors will eliminate rational indeterminacy.

[36] Recall that I have loosely defined cooperation as exchanges or trades between very few people, *supra* at 4.

Nonetheless, I start out with a short analysis of cooperation norms (1. below). I do this in spite of theorists' confidence because cooperation too is plagued by rational indeterminacy and thus faces the problem of norm stability. Once it is understood how precarious cooperative reputation is, and why it nevertheless exists, it turns out that other norms can build on the same mechanism (2.). The overall model so developed is then compared to two other norm theories based on rational choice, namely Richard McAdams' esteem model (3.) and Eric Posner's signaling model (4.).

1. COOPERATION NORMS

I have noted in the first section that even cooperation between pairs of players is much more troublesome than is usually perceived.[37] Reputation norms enforcing cooperative behavior are even more puzzling: How can a player be induced to sanction a wrong done to *another*?[38]

Conventional analysis of reputation very often relies on private information.[39] This way of looking at reputation is characterized by assuming a piece of *model-exogenous information*, i.e., information that is not itself explained as a result of the theory. A typical example is that some players command a better production

---

[37] Cf. *supra* at 6.

[38] Thus, I leave aside the problem of simple 2-players iterated Prisoners' Dilemmas. One solution to this problem is that conditional play in the 2-players iterated Prisoners' Dilemma rests on the threat of earning a reputation as a wimp.

[39] So do the seminal papers David M. Kreps & Robert Wilson, Reputation and Imperfect Information, 27 J. Econ. Th. 253 (1982) and David M. Kreps *et al.*, Rational Cooperation in the Finitely Repeated Prisoners' Dilemma, 27 J. Econ. Th. 245 (1982).

technology making cooperation less costly. If there is a market, these players can offer better prices under which other, less well-equipped players cannot supply the same quality. The reputation mechanism then works to sort out the less well-equipped players by the low quality they deliver at the low market price. Here, reputation is a revelation of private information (production technology) through observable behavior (quality delivered in the past). Note that the differences in production technology are *assumed*, not *explained*. Therefore, it is fair to say that these models are based on model-exogenous information. What those models do explain is how reputation emerges from such a given difference.

Reputation norms based on model-exogenous information are inevitably stable: It is always credible to promote such a norm because the good sellers as well as the buyers only benefit from adhering to it. Conversely, opposing the norm is self-defeating because it only shows that one does not benefit, which means that one is a bad seller. Therefore, a theory of reputation based on model-exogenous information does not need the complicated theories of norm stability laid out in the previous section. In particular, reputation based on model-exogenous information does not rely on network effects: If a player can fully observe the history of the game she need not bother if others apply the same reputation norm. The reason is that here the norm is nothing more than a Bayesian inference to estimate the probability of a seller having the good technology.[40]

[40] Note already that reputation based on model-exogenous information closely resembles signaling theory. Cf. III.4. at 43 *infra* for E. Posner's signaling model. At closer inspection, there is no

However, that is not yet the full story. There is a second type of reputation that, in a way, *creates* information. Greif's analysis of the Maghribi Traders' reputation norm is an example of such a model that endogenously explains the relevant information, rather than assuming it. This difference clearly emerges in the language of the above quote: According to Greif, cheaters would be punished "*despite* the fact that cheating in the past does not indicate that the agent is a 'lemon.'"[41] The reputation theory proposed in this paper is concerned with this second type of reputation, *not* with the first one based on model-exogenous information. Therefore, to simplify terminology when I speak of "reputation" I refer only to this second type (unless otherwise indicated).

We have seen that reputation norms of the first type (i.e., reputation based on model exogenous information) will always be robust as it is always unambiguously best for players to stick to the norm. The puzzle mentioned above of why players would punish a wrong done to another player pertains only to second-type reputation norms. It is at this point that the concept of norm stability developed in the previous section comes into play. The key to stability of (second-type) reputation norms is their strong network effects. Recall that, in the case of a stigmatized Maghribi trader-agent, principals had to pay a higher remuneration in a long-term cooperative

---

difference at all: Delivering good quality to other players is a signal that one is a good seller, just like boasting expensive premises may be a signal that one is expecting to stay in business for long (and thus will honor one's obligations).

[41] Cf. *supra* at 14 and Avner Greif, Contract Enforceability and Economic Institutions in Early Trade: The Maghribi Traders' Coalition, 83 Am. Econ. Rev. 525, 534-5 (1993), (italics in the original).

relationship to ensure faithful performance. It followed that cheaters were more expensive agents so that principals were driven to shun them by turning to cheaper supply of the service. Thus, in contrast to reputation based on model-exogenous information, the competitive edge of faithful agents here is generated endogenously by the norm itself. A Maghribi trader is well advised to mistrust a cheating member and turn to a less suspicious fellow as long as he is confident that the norm will remain effective. But an attempt to change the norm also fails to be preference compatible. The incentive to push for changing the reputation norm is in most cases purely opportunistic.[42] That is, only an agent who has cheated has an incentive to get rid of the norm while honest agents and principals have a preference for retaining it. It follows that the incentive to amend the norm is confined to the individual cheater (plus possibly a number of players who maintain long-term cooperative relationships with him). The vast majority of players is interested in sustaining the norm and therefore opposed to undermining its credibility. Therefore, amending the norm fails to be compatible with the preferences of the vast majority. [43]

2. Beyond Cooperation: Internal Loss Norms

We have seen why cooperative reputation norms can be quite robust. Cooperative reputation induces players to forego the potential gains from cheating, thus benefiting cooperation partners. Yet there are other norms that are just as costly to

[42] It may not be opportunistic if there is a probability that players are falsely accused of cheating.

[43] Cf. *supra* at 21.

comply with while not delivering (immediate) benefits to cooperation partners. These norms are of particular interest because they seem to be most at odds with rational choice. Therefore, the analysis of such norms in this subsection goes at some length to introduce the concept of *internal loss norms* (a)), address the key question of how these norms accomplish stability (b)), and relate them to the problem of collective goods (c)).

a) Internal Loss Norms

To get a better sense of the problem, consider the example of an environmentalist: Say I insist that my friends pay regard to preserving nature and the environment. If I am an attractive friend to have, then perhaps some people may be induced to rituals such as separating garbage. Yet the problem is that, by pressuring my friends, I cannot noticeably improve the environment. At the same time, separating garbage and similar activities are not too much fun, and they cost leisure. Chances are that my friends would be fond of me even more if I allowed them to maintain our relationship without caring for the environment. Hence, both my friends and I would be better off if I gave up my obsession with the environment. So why do we not spare us the hassle and stop feeling responsible for things that we cannot change anyway?

More specifically, my demand to protect the environment faces a pareto-superior equilibrium viewed from my standpoint and that of my friends: The environmentalist norm leads us to give up value from our cooperative relationship. Since norms like the environmentalist one create a loss from the point of view of

cooperation partners, I refer to them as *internal loss norms (ILN)*. By contrast, the cooperation norms considered in the previous subsection reduced cheating between the cooperation partners; they generated an *internal gain*.

In principle, the cost imposed by ILN may just be lost. Alternatively, the costly activity required by the ILN can have a positive or negative effect on the group and on outsiders. It can consist of promoting a collective good, as in the example of the environmentalist norm. It follows that ILN can contribute to solving collective good problems (which I will discuss below c)). Note, however, that the definition of ILN only requires that players incur an internal loss.

b) Internal Loss Norms and Stability

The definition of ILN already points at the very problem they pose: How can a norm be stable if the players immediately involved could, in principle, do much better? Why would they not maximize their joint benefit by agreeing to abolish the ILN?

The trouble with ILN is that the players directly involved have every incentive to abolish the norm. Hence, ILN face a particularly adverse problem of norm stability. Under the norm stability theory developed in the previous section it is clear that they need to call in powerful network effects, and that they must be preference compatible. An ILN can quite easily be preference compatible, particularly if it helps supply a collective good. The more serious problem is to arrange strong network effects. ILN themselves lack network effects: While they may provide collective

goods and, to the extent they do so, increase every individual player's welfare, one does not have to comply with the ILN oneself in order to participate in the collective good. On the contrary, not complying with an ILN is, by the definition of ILN, the dominant strategy choice. ILN seem to thwart the only mechanism that could give them stability.

If ILN cannot win on their own they must look for a powerful ally. Such a potential ally exists – it is the cooperative reputation norms analyzed in the previous subsection. Cooperative reputation norms boast strong network effects because self-interest induces players to avoid reputational outcasts.[44] The solution for ILN, therefore, is to link themselves to a cooperative reputation norm, which by its very nature has a large network with strong network effects.[45] At first look, the idea of linking norms may seem dubious. Also, it seems odd that a cooperative norm can be harnessed to impose behavior that "wastes" cooperative gain. But linking particular norms is itself a norm. Its stability follows the general conditions specified for norm stability, namely network effects and preference compatibility. Linking an ILN to a cooperative reputation norm simply redefines reputation and can as such be founded on network effects and preference compatibility. As we have seen, standard cooperative reputation is actually no different: From the point of view of one particular cooperative relationship, the partner's reliability in *another* relationships is equally a

[44] See *supra* 1. at 23.

[45] It is in the nature of the theory advocated here that this link does not flow from model-exogenous information (cf. *supra* 1. at 23); for a model of ILN based on exogenous information see Eric Rasmusen, Stigma and Self-Fulfilling Expectations of Criminality, 39 J. L. & Econ. 519 (1996).

"waste". Nonetheless, reputation links performing under one's obligations in one relationship to another cooperative relationship. If reputation can do a job as precarious as this, then it can also enforce ILN.

To illustrate, consider once again the Maghribi community. Presumably, being a reputable member of this community required professing the Jewish creed. On the other hand, retaining the Jewish religion was probably a costly behavior at times; it was, in other words, an ILN. Of course, membership awarded certain benefits, such as participation in the cooperative reputation norm, which compensated for the cost. Nonetheless, the key problem of ILN arose for the Maghribi community as well, namely why players should not take the benefits of cooperation while breaking with the ILN. The solution of the community, most probably, was to link a trader's reputation to, amongst other things, not abandoning his religion. If one of the traders had wished to give up his religion he would have lost his reputation and thus the benefits of being hired as an agent. To avoid this loss he would have had to convince the whole community of amending the norm, which would have been extremely costly (if at all possible).

It can be retained as a result that ILN gain stability when they link themselves to cooperative reputation norms. Yet so far the argument has only been that such linking is possible at all. It is a different question how powerful this explanation is. Basically, the ability of cooperative reputation to sustain ILN depends on the value of cooperative reputation for the individual player. The question is if reputation is strong enough a basis to explain the variety and total cost of the ILN in the real world, or at

least a significant part of them. The claim of a reputation theory of norms, naturally, is that reputation generates a lot of value for players and can extract considerable value for sustaining ILN. There are (at least) three points to be made for that claim.

The first point is a defense against a potential objection. For one might argue that we observe most individuals to trade with a relatively small and constant set of partners. This seems to suggest that long-term cooperative relationships predominate and that reputation is not too important after all. The underlying assumption is that reputation and long-term relationships are substitutes. This assumption, however, is only partly correct. It fails to realize that long-term cooperative relationships and reputation are also complements. At its heart, cooperative reputation consists of increasing the cost of cheating: The cheater loses not only one relation but *any* possible relation. Of course, if players did not have an interest in preserving outside options, reputation would not have much weight. However, I claim they do have such an interest even in spite of considerable investment in present relations and in spite of little actual switching.[46] This interest consists of the benefits of (potential) competition: The option to switch partners is both a check against opportunism and an incentive for partners to enhance their attractiveness. Hence, without the option to leave, even a long-term relationship loses a great deal of its value. Consider marriage, a relationship that is meant to last a lifetime. It seems obvious to me that a spouse

[46] Note that without such an interest, reputation and the benefits of enhanced enforcement would not exist.

loses a lot of leverage within marriage when she gives up all her outside options (ranging from career opportunities to personal attractiveness as friend or companion).

The second reason to think that cooperative reputation is powerful pertains to specific types of cooperation. Often the very value of a relationship is gaining access to a relationship network or group. Whenever exchanging social links is part of the deal – and it probably is quite often – one's reputation directly affects the quality of the good that one offers. For instance, in marriage the social relations and status of a spouse is a crucial asset for exchange. Similarly, if a company hires a politician retiree it does so in order to gain access to the public (other politicians, government officials, journalists etc.). If the politician is prosecuted for tax evasion she may lose her reputation with the group that she was supposed to provide access to.

There is a third point bolstering the claim that reputation is strong enough to explain the abundance of ILN that we observe around us. The tying of ILN to cooperative reputation, in principle, solves the problem of why people would punish someone who violates an ILN.[47] The corresponding incentives are brought about by the self-enforcing character of cooperative reputation, that is, by the fact that it is more expensive to hire a player with a bad reputation (if the job requires trustworthiness). While this is the basic mechanism, it can give rise to a second mechanism that extends the power of ILN. This second mechanism consists of

[47] In the context of ILN that supply a collective good, this problem is often referred to as "second order collective action problem". See Robert C. Ellickson, The Market for Social Norms, 3 Am. L. & Econ. Rev. 1, 18 *et seq.* (2001) and the reference to Buchanan in n. 12 *supra*.

imposing a loss in reputation on anybody who fails to punish (e.g. still cooperates with) the transgressor. In other words, the second mechanism is itself an ILN that requires participating in the enforcement of another norm.[48] As a consequence, partners will urge each other to abide by the norm in order to spare themselves the conflict between losing their reputation and having to terminate the relationship. This way, enforcement of ILN can reach an additional fraction of players who would otherwise prefer a loss in reputation to incurring the cost of compliance.

It emerges from the three arguments that cooperative reputation can be quite a powerful vehicle for ILN. Generally, reputation is an additional device to check opportunism and to enhance cooperative opportunities. If players have such an additional option at their disposal they will probably use it. ILN can extract a considerable part of the value of cooperative reputation – in fact, it can extract *all* the value. When this happens individual players cannot pull back from the reputation mechanism any more as this would mean giving up the cooperative opportunities from reputation, which at least compensate for the ILN. Thus, players are trapped in the collective action problem of getting rid of reputation. But much unlike other collective action problems, this one can be beneficial: It can help to overcome other, socially harmful collective action problems.

---

[48] This is so because the norm is shaped by what outsiders want, and they will often prefer the norm to have a strong effect. See c) *infra*. – The argument has a similar structure as E. Posner's notion of sanctioning as a signal (see Eric A. Posner, Law and Social Norms 25-6 (2000)) and McAdams' "secondary enforcement norms" (see Richard H. McAdams, The Origin, Development, and Regulation of Norms, 96 Mich. L. Rev. 338, 372-5 (1997)).

But before reaping this result of our analysis, another condition for the enforcement of ILN deserves note, namely *observability of compliance*. Observability is indispensable for a reputation account of ILN: Only observable behavior can be made subject to reputational sanctions. A less trivial point is the effect that reputation norms have on the sharing of observations among members of a norm community. In particular, the reputation mechanism and its network effects help explain why people like to gossip about others' failure to comply with social norms: Players have a strong interest in obtaining information so as not to deal with a player who h as lost her reputation and therefore is unreliable.

c) Internal Loss Norms and Collective Goods

A case has been made, in the preceding paragraphs, for the proposition that a substantial amount of ILN can be sustained by tying the ILN to cooperation norms; a person violating the ILN suffers a loss in her cooperative reputation. Now, this result is to be applied to the kind of behavior that has always troubled norms scholars most, the puzzle of social norms that overcome the collective action problem and produce collective goods: Collective good norms induce people to vote, to express disapproval of others' littering, to pay taxes in spite of manifold occasions to evade them, and to volunteer for honorary service (or even for the army in times of war). In all these instances, people incur losses – sometimes very small costs, but quite substantial costs on other occasions. If, as I have argued, the reputation theory of norms can explain the stability of ILN then it has the key to this part of the puzzle: Collective good norms *are* ILN. At the same time, they are a peculiar type of ILN: Not only do they impose *internal* losses on the partners in cooperation but also do they produce

external benefit in the form of a collective good. While the former element has been dealt with, the latter requires further analysis. The question is this: Given that ILN can be sustained, why would they tend to create an external good?

Reconsidering the conditions of norm stability worked out in the previous section it appears that the analysis of ILN has exploited the first condition – strong network effects – but not the second one, preference compatibility. This second condition provides an answer to why ILN tend to produce collective goods. Recall again the way ILN use cooperative reputation to make players incur cost: The behavior governed by ILN must be observed by other players. Also, at least a number of the observers must be outsiders, that is, they must learn of the behavior although they do not stand in a close cooperative relationship with the actor. It is these outsiders who decide on the acting player's reputation by judging the action as norm compliance or violation. While partners in a cooperative relationship try to maximize the internal benefit, outsiders neither win nor lose from the individual player. Their only choice is between preserving the ILN and amending it or, as the case may be, between inventing or not inventing a new ILN. It follows that, in principle, outsiders, as decision-makers, are interested only in whether or not the norm should *generally* exist and not in the implications of that decision on the particular violator. Yet outside players are induced not only to abstract from the individual violator. Also, they are forced to abstract from their own, purely selfish ends: Reputation norms only work if they are applied uniformly by a relevant part of the population. Hence, a reputation norm can only exist if it is widely accepted, and such widespread acceptance requires that the norm favor many members of the population (preference compatibility).

Therefore, ILN tend to be impartial towards both the players *judged* and the individual players *making the judgment*. Effectiveness of the norm requires consensus, consensus requires the norm to produce a collective good.[49]

In sum, the idea is that norms – and ILN in particular – are shaped by unaffected outsiders who have the freedom to decide according to their true preferences over norms *in general*. ILN tend to supply collective goods because most players prefer collective goods to be produced even if that means that all players (including themselves!) have to contribute. Note that this argument is paralleled by the role of public opinion in political theory and by the law's obsession to expose powerful actors to the public.[50] The intuition there is that public speech constrains the speaker's selfishess. [51] This intuition probably is correct for a similar reason as the one advanced here: Defying a powerful player is an extremely costly undertaking for other players. Therefore, as long as a powerful player acts within a small group she may stretch the rules in her favor. But even the most powerful individual cannot

---

[49] The argument in the text is the same as the one adduced above for the preference compatibility condition, cf. *supra* II.2. at 18

[50] Straightforward examples are the right to a public trial and the publicity requirements for the executive and legislative branch. It is a related point that people often behave differently in their capacity as consumers and citizens. In the latter role, they are often more attentive to the public good, Cass R. Sunstein, Social Norms and Social Roles, 96 Col. L. Rev. 903, 923-4 (1996). Still another application is the competitive process of science (and the risk of its being corrupted by political preferences), see Ronald H. Coase, How Should Economists Choose?, in Coase, ed, Essays on Economics and Economists 15, 30-1 (1995).

[51] Cf. Joseph Farrell & Robert Gibbons, Cheap Talk with Two Audiences, 79 Am. Econ. Rev. 1214 (1989).

control how an audience of thousands or millions form their opinion. Therefore, it is on the public stage that restrictions can be imposed on powerful actors.

Another remark is in point. The idea that ILN are shaped and enforced by outsiders also has implications for the problem of observability. In the case of ILN, if the norm is violated there is no individual victim. On the contrary, cooperation partners who may have witnessed the violation have an incentive to acquiesce to a deviation, split the gain, and not tell anybody.[52] Thus, an ILN has a stable design only if players outside a long-term relationship can observe norm compliance, or if the norm provides incentives to disclose norm violations and if these incentives outweigh the gains from transgression. In order to meet this requirement, ILN sometimes prescribe visible behavior, such as a youth gang's courage tests or the religious obligation of attending public worship. However, with some ILN compliance is not publicly observable. In these cases, players who deviate from the norm experience the fear that their violation is revealed. That is, they experience shame. Notably, shame is not confined to the perpetrator. Family relatives and other long-term partners are also reluctant to expose a perpetrator's violation. Only when the violation is publicized

[52] One difficulty with such a hidden deviation is that the norm violator makes herself subject to holdout. On the other hand, opportunism among the conspirators is alleviated if the norm sanctions acquiescence to the transgression as well or if the conspirators maintain a long-term relationship under which both partners violate the norm.

they detach themselves from the perpetrator if they can afford to.[53] Family members often cannot. Hence they are anxious to avoid disclosure of violations.[54]

Obviously, the picture painted so far remains coarse. In the context of the present article, it is impossible to provide anything more than a rough sketch. The optimistic message, nevertheless, is that ILN drift towards supplying collective goods.

3. COMPARING MODELS: MCADAMS' ESTEEM THEORY OF NORMS

By now, the main elements of the reputation theory of social norms have been presented. The remainder of this section aims at enriching the theory further, particularly by putting reputation theory in perspective with two prominent pieces of norms scholarship. In this subsection, reputation theory is related to Richard McAdams' esteem theory of norms.[55]

The core assumption of esteem theory is that people have a slight preference for something that other people can give or withhold at zero cost: esteem. McAdams himself emphasizes that the assumption serves to eschew the collective action

---

[53] E. Posner reports that many of Oscar Wilde's friends were aware of his sexual orientation but started to shun him only when it was made public, see Eric A. Posner, Law and Social Norms 24 (2000).

[54] Tragically, they even avoid publicity when they themselves are affected by a norm violation, such as family violence and sexual abuse.

[55] See Richard H. McAdams, The Origin, Development, and Regulation of Norms, 96 Mich. L. Rev. 338, 355 *et seq.* (1997).

problem of norm enforcement:[56] Because esteem is costless it is not subject to a free rider problem.[57] McAdams exploits this assumption to deduce a number of very interesting results: For example, although the preference for esteem is assumed to be slight, McAdams dexterously uses it to derive even very costly norm-guided behavior.

My point here is not to criticize those results. Instead, what worries me is that esteem has the flavor of an *ad hoc* assumption. McAdams is certainly right to insist that esteem is a ubiquitous social fact.[58] Such a fact can figure as an assumption for the theory. However, it can as well be the *explanandum* of such theory. In what follows, I follow the latter path and explain esteem as an upshot of the reputation theory presented here. Supplying a model to McAdams' assumptions avoids some objections against the esteem model: For instance, if esteem was costless to award, why would not players trade it in for private benefits? That is, why would not esteem become an ordinary good – to the ultimate effect that players would ascribe it an opportunity cost?[59] To invalidate objections like these, it may be helpful to understand esteem as reputation. With this interpretation, McAdams' theory is quite compatible with the one set out here, which allows combining their results.

---

[56] Cf. n. 47 *supra*.

[57] See Richard H. McAdams, The Origin, Development, and Regulation of Norms, 96 Mich. L. Rev. 338, 352 *et seq.* (1997).

[58] See Richard H. McAdams, The Origin, Development, and Regulation of Norms, 96 Mich. L. Rev. 338, 355 (1997) and Richard H. McAdams, Signaling Discount Rates: Law, Norms, and Economic Methodology, 110 Y. L. J. 625, 681 *et seq.* (2001).

To see that esteem can be identified with reputation it is important to clarify the concept of reputation. Most importantly, reputation is not necessarily an all-or-nothing sanction. Consider again the Maghribi traders. Recall that a stigmatized trader was not necessarily out of business altogether but simply cost more to his principals. This is a hint that reputation can be seen as a matter of degree. There may have been cheating of different severity resulting in different degrees of loss of reputation. On markets, we are used to fine differentials in the reputation of a firm, brand, product, or professional with nuances ranging from the world-renowned surgery specialist to the quack. Such continuous reputation extends beyond markets. For instance, in spite of the dualistic character of moral norms people measure moral character in degrees, rather than using two disparate categories of "the good" and "the evil."

With this idea of continuous reputation in mind, McAdams' notion of esteem can simply be set in one with reputation as construed by reputation theory. Just as McAdams emphasizes with respect to esteem, continuous reputation is a relative measure. That is, one aims at earning more reputation/esteem *in comparison to others*.[59] If most people do not care for the welfare of the community as a whole then being trustworthy in bilateral trading already earns one an average reputation.[60] By contrast, in a society of saints an average reputation is associated with a great deal of

---

[59] Also, one wonders if more esteem from player *A* could substitute for player *B's* esteem, and if so, at what rate of substitution.

[60] See Richard H. McAdams, The Origin, Development, and Regulation of Norms, 96 Mich. L. Rev. 338, 357 (1997).

self-denial, i.e. high costs through ILN. Also, like McAdams' esteem, continuous reputation can be withheld at no cost. In my analysis above I have emphasized that only outsiders who do not incur much risk can enforce ILN.[61] I presume that McAdams would accept the same *caveat* for esteem: Players maintaining a valuable cooperative relationship will be reluctant to withhold reputation/esteem. Thus, the concepts of reputation and esteem are well compatible in this respect, too. Finally, there is a correspondence to another one of McAdams' assumptions: McAdams observes that a single individual's esteem is worth little but that esteem from many individuals adds up to a noticeable effect.[62] The same is true for reputation by virtue of network effects: I may be perfectly happy if some single individual thinks I have lost my reputation (or should have lost it because of what I have done). The situation becomes worrying only if many people disrespect me so that the alleged loss of reputation becomes self-enforcing.

---

[61] Cf. *supra* III.2.c) at 34 Obviously, there is a cost in excluding a player from future cooperation (cf. the Kreps reference in n. 35 *supra*). But if the reputation mechanism works smoothly, the cost is spread on the group as a whole. A single member of the group cannot do anything to avoid this cost. She cannot, in other words, award reputation to other players independently from the social norm in effect. It is for this reason that players cannot promise to award reputation in exchange for a side-payment – which was one of the problems with McAdams' esteem assumption, see *supra* at 40 text accompanying n. 59.

[62] See Richard H. McAdams, The Origin, Development, and Regulation of Norms, 96 Mich. L. Rev. 338, 365 (1997).

Reputation theory is so close to McAdams' esteem theory that his results apply in a reputation framework, too.[63] These results owe their strength to the insight that esteem is a relative value, i.e. that one aims at being esteemed *in comparison to others*. I have just demonstrated that the same is true for (continuous) reputation. One very important implication is that sanctions increase as the number of perpetrators drops. The reason is that the number of norm violators shrinks and thus the number of players ranked relatively higher soars. Correspondingly, norms can exhibit self-amplification up to the point where compliance has become the standard behavior and nobody gains any more *positive* reputation from complying (but merely avoids reputational harm). These dynamics cannot only crunch a minority of deviators but they can also cause new norms to rise. "Heroes" may trigger the emergence of a new norm by incurring sacrifices. So long as their endeavor is that of a few dispersed individuals their sacrifice remains futile. If they fail they are ridiculed. However, if they succeed they trigger a competition for compliance up to the point where the proposal has become the new standard. One crucial condition of success for a hero is to be a "norm entrepreneur" at the same time, that is, to convey to their audience why the new behavior is useful and thus deserves esteem (or reputation). Hence the mixture of eloquence and resoluteness that accounts for charismatic leadership.

While reputation theory greatly benefits from esteem theory it is, on the other hand, more comprehensive than esteem theory. The most important point certainly is

---

[63] See Richard H. McAdams, The Origin, Development, and Regulation of Norms, 96 Mich. L. Rev. 338, 365 *et seq.* (1997).

that reputation theory explains rather than assumes reputation/esteem. Yet this is not the only point where reputation theory can stretch reductionism further.[64] Another is gossip. As other players are eager to keep track of a player's reputation there is no need to premise gossip on an intrinsic pleasure, as McAdams does.[65] Gossiping is nothing else than an exchange of information needed to adjust the amount of trust one can place into others. Similarly, reputation theory explains the consensus that McAdams requires as a condition of a norm.[66]

To conclude, McAdams' norms theory based on esteem adds important insights for reputation theory, while reputation theory may help identify the forces driving reputation/esteem. In explaining norms, the two theories can be considered as complements rather than substitutes.

4. Comparing Models: E. Posner's Signaling Theory of Norms

Eric Posner has worked out an impressive theory of norms based on the idea of costly signaling.[67] Like McAdams, Posner follows the rational choice path in ana-

---

[64] However, for McAdams' criticism of excessive reductionism see the reference in n. 7 *supra*.

[65] See Richard H. McAdams, The Origin, Development, and Regulation of Norms, 96 Mich. L. Rev. 338, 362 (1997) as opposed to *supra* at 34

[66] See Richard H. McAdams, The Origin, Development, and Regulation of Norms, 96 Mich. L. Rev. 338, 358-9 (1997). – Under reputation theory, a uniform norm is an implication of the network effects of reputation norms, cf. II.1. and II.2. *supra* at 14 *et seq.*

[67] See only Eric A. Posner, Law and Social Norms (2000).

lyzing social norms.[68] However, while McAdams' model is a complement to reputation theory, Posner's theory is a substitute.

Generally speaking, signaling is a costly behavior that conveys some piece of model exogenous private information in a credible way. Because it relies on model-exogenous information it stands in the same camp as reputation based on model-exogenous information that I have discussed above (reputation of the first type).[69] Like in the case of reputation based on model-exogenous information, the signaling player knows a particular variable that only she can observe. Other players would like to cooperate with the sender only if the variable has the "good" value; otherwise, the other players prefer not to cooperate. Consider again the hypothetical of the sellers with the good and the bad production technology.[70] As we have seen, good sellers can demonstrate their superior technology by delivering high quality to all customers. This would commonly be called reputation (based on model-exogenous information). By contrast, a *signal* might consist of offering one delivery for free in order to demonstrate quality, or of making a binding promise to take back any good that a customer may wish to return.

Posner uses the signaling framework to explain social norms. In doing so, he has to choose which model-exogenous information to assume for the model, i.e. the information conveyed by the signal. For this purpose, Posner chooses players'

---

[68] In a way, he is even more a rational choice scholar than McAdams because he attempts (as I do) to treat social norms without resorting to normative motivations.

[69] See *supra* 1. at 23.

preferences with respect to consumption over time, i.e. their discount rates.[71] A person with a high discount rate prefers present consumption to future consumption relatively more than a person with a low discount rate. Consequently, a person with a high discount rate is less trustworthy because she values the loss of a relationship or a punishment in the future relatively less than a person with a low discount rate. Thus, Posner distinguishes bad cooperative types (high discount rate) from good cooperative types (low discount rate). With that assumption in place, the signaling framework is ready to be applied on the subject of norms: Behavior guided by social norms is, according to Posner, costly and visible signaling that is intended to demonstrate the sender's good type.[72]

Posner's signaling approach is especially attractive when it is compared to the difficulties that reputation theory had in coping with ILN. As we have seen, ILN must prevent the players from abolishing the ILN, which would enable them to retain more value for themselves. Posner does not explicitly address norm stability – simply because he does not have to. He can afford to ignore the problem because the signaling

---

[70] See *supra* at 23.

[71] See Eric A. Posner, Law and Social Norms 18-9 (2000). In a paper with Professor Goldsmith on international relations, Posner offers "political stability" as a private information to be signaled. See Jack L. Goldsmith & Eric A. Posner, Moral and Legal Rhetoric in International Relations: A Rational Choice Perspective, John M. Olin Law & Economics Working Papers (2d series) No 108, 8 (Chicago, 2000).

[72] More precisely, a norm is "a description of the behavior that emerges in … signaling equilibriums." The latter definition includes cheap behavior that does not serve as a signal but avoids being punished by others who signal by punishing deviations, see Eric A. Posner, Law and Social Norms 24-6 (2000).

approach is immune to it: Players cannot agree to abolish a signaling ILN because they need signaling to overcome a problem of asymmetric information.

This short summary of Posner's signaling model highlights why it must be a substitute, not a complement to the reputation theory of norms. Reputation in the narrow sense I use the term here is *not* based on model-exogenous information; instead, it is a social construct entirely.[73] Posner's signaling model is on the opposite side. Therefore, the two approaches substitute for each other. Posner's model is better suited whenever we think that there is a relevant piece of private information, which cooperative partners need to know and which can, at the same time, be conveyed in a credible way by means of a costly signal. Both conditions must be met for the signaling account to work. If they are not then reputation theory is the only option.

It is at this point that the critique against signaling theory should be taken into account.[74] One argument strikes me as particularly important. I have said that the signaling model must assume model-exogenous information that is at the same time relevant for cooperation and suitable to be signaled. This is already a significant

---

[73] As has been discussed at some length *supra* under III.1. at 23

[74] A lot of the critique stems from Posner's decision to rely on one single piece of information – the discount rate – as the signal's message. See Richard H. McAdams, Signaling Discount Rates: Law, Norms, and Economic Methodology, 110 Y. L. J. 625, 654-663 (2001) and Paul G. Mahoney, Norms and Signals: Some Skeptical Observations, 36 U. Rich. L. Rev. 387 (2002) (suggesting that the less well-off would have a greater need to show their reliability). Yet it seems Posner would be willing to accept a different interpretation, cf. n. 71 *supra*. Under such an understanding, the work of other authors can also be ascribed to the signaling approach to social norms. For instance, not committing

limitation. But even if these conditions obtain there must, incrementally, not be a cheaper alternative to signaling in order to credibly convey information.[75] One obvious candidate is a cooperative reputation norm.[76] To cite one of Posner's examples, if people's donating to the opera is a signal that serves to facilitate *cooperation*, then it must be impossible to achieve the same result through *cooperative* reputation. If the donator wants to make private friends it would be better for her to show how well she cooperates with the friends she already has because, that way, she would not only signal but at the same time enhance the value of her existing friendships – instead wasting money on the opera. Of course, cooperative reputation may not be an option. It may be that cooperative behavior is not observable well enough, or a newcomer may not yet have cooperative relations.[77] One implication is that we would expect signaling at an initial stage, say, when entering a new market. Later, we would expect it to give way to cooperative reputation.[78]

---

crimes can be seen as a signal of high productivity. See Eric Rasmusen, Stigma and Self-Fulfilling Expectations of Criminality, 39 J. L. & Econ. 519 (1996).

[75] See Richard H. McAdams, Signaling Discount Rates: Law, Norms, and Economic Methodology, 110 Y. L. J. 625, 677 (2001).

[76] See Richard H. McAdams, Signaling Discount Rates: Law, Norms, and Economic Methodology, 110 Y. L. J. 625, 674-676 (2001).

[77] A donator may seek for a public role and thus be interested in the broad visibility of a donation. Yet this hardly explains donations from all those with a more narrow audience.

[78] See Paul G. Mahoney, Norms and Signals: Some Skeptical Observations, 36 U. Rich. L. Rev. 387, 391-2 (2002) (for small groups). – Of course, Posner anticipates this objection. He responds by pointing to weak memories, insufficient flow of information, and changing outside conditions, see Eric A. Posner, Law and Social Norms 20-1 (2000). The problem I see with this argument is that signaling becomes especially expensive if it is used persistently. Hence, there is a strong incentive to come up with an alternative to persistent signaling. To be sure, a reputation driven ILN is equally expensive.

Another cheaper substitute for signaling can be private or public memory. Provided that the information conveyed by the signal does not change, it seems cheaper to store that information, be it in one's head or through gossip.[79] This substitute should be available at least for small and close groups where information flow is good.

The point is that signaling has to rule out alternatives, and that its applicability will be limited. Note that, by contrast, reputation theory confronts norm stability directly. Therefore, if a reputation ILN can be shown to be stable because of network effects and preference compatibility, it cannot be contracted around. By the same token, it need not compete with other arrangements that may be more attractive for the cooperative partners.

IV. Examples

So far, the reputation theory of social norms has been presented on a theoretical level. To vivify the picture, a few examples may be helpful. In the context of this exploratory paper, I will not take such "hypotheses testing" very far but some of it

---

But the reputation theory proposed here is backed up by an account of norm stability, which constrains players' ability to get rid of it.

[79] Mahoney perceives a somewhat different problem for signaling theory. In his view the fact that information conveyed by signals is valuable should lead to information trading. From the fact that such trading rarely takes place he concludes that signaling cannot be as important as Posner believes. See Paul G. Mahoney, Norms and Signals: Some Skeptical Observations, 36 U. Rich. L. Rev. 387, 394-6 (2002).

will be helpful to make the point. The examples offered aim at providing an intuition of how the reputation theory of norms plays out in concrete analysis.

1. Private Provision of Collective Goods: Donating and Volunteering

The first example of a norm that I consider through the lens of reputation theory is donating and volunteering. It is astonishing from a rational choice perspective that people donate or volunteer for helping the poor, for protecting the environment, for improving education, and for providing other collective goods. The challenge is exacerbated by the fact that donating and volunteering is an expensive activity. It cannot be played down to a minor defect of rationality resulting in some negligible loss.

Of course, "voluntary" contributions to collective goods are precisely what the concept of ILN is designed to capture. Thus, explaining those oddities (from a rational choice perspective) does not demand more than applying the framework of stable reputation ILN. However, the intuition behind this solution may be somewhat hard to grasp. For instance, we rarely find a norm requiring to shun people who do not donate a certain standard portion of their income. But as we have seen, reputation must be thought of as a continuous variable.[80] Accordingly, a reputation norm is not a rigid threshold that either awards or withholds reputation. Rather, a reputation norm should be conceived of as a function assigning a reputation value to a multi-

[80] As I argued earlier, reputation can explain most of the properties that McAdams ascribes to his esteem assumption. Cf. *supra* III.3. at 38

dimensional vector of activities. That function determines whether activities are substitutes for or complements to each other. Activities like donating or volunteering are complements to a bunch of more basic behavioral patterns such as refraining from criminal activity, from being polite to others, not cheating on them, etc. It certainly does not pay in terms of reputation to donate generously to the local museum while being caught stealing in the supermarket. Conversely, the various possibilities for contributing to collective goods are substitutes. It does not matter so much, as far as general reputation is concerned,[81] whether you donate to the museum or to Greenpeace. Probably, you even do not need to donate at all as long as you contribute in some other way, e.g. by participating in the political process or by showing interest in cultural developments. However – and this is the main proposition of reputation theory – a person without any such commitment to the common good will be less esteemed, considered boring, less valued as a friend or conversation partner, less attractive on the marriage market, and so forth.

So far, the story told by reputation theory does not sound different from that of signaling theory. In Posner's view, donating serves people as a signal of being a cooperative type.[82] I have already mentioned one important limitation of signaling, namely that it should often be defeated by cooperative reputation, especially when

[81] I speak of *general* reputation to distinguish it from reputation with a particular group. Of course, a benefactor of the local museum will enjoy a particular reputation among those interested in the arts.

[82] See Eric A. Posner, Law and Social Norms 50 *et seq.* and 65-67 (2000).

one already has many cooperative ties.[83] By contrast, a reputation norm can persistently extract expenditures and effort from the player; they have to comply or else they lose their reputation.

Also, reputation theory yields additional predictions compared to signaling theory (making it both richer and more susceptible to falsification). At first sight it seems that a signal need only be costly.[84] Still, Posner realizes that the institutions funded by private donations perform particular services, and that signaling theory would remain incomplete without explaining why, apparently, burning cash does not substitute for donating. Posner responds by introducing another requirement for a signal, namely that it attract attention. It is for this reason, Posner argues, that money is given to attractive and spectacular projects.

I do not doubt that this point is highly relevant. Visibility and attracting attention is important for a reputation account, too.[85] Still, it falls short of explaining many forms of giving that seem not to maximize visibility. It is telling that Posner uses the opera as an illustration instead of, say, charity.[86] While the former is entertaining and can be seen as satisfying intrinsic preferences, the latter is beneficial for society but not particularly interesting for the audience it aims at: People devote time and money

---

[83] *Supra* at 48.

[84] At one point, Posner posits that "*any* costly action can be a signal" (Eric A. Posner, Law and Social Norms 24 (2000), emphasis added). Cf. Richard H. McAdams, Signaling Discount Rates: Law, Norms, and Economic Methodology, 110 Y. L. J. 625, 640 (2001).

[85] On observability as a precondition of enforcement see *supra* at 34.

[86] See Eric A. Posner, Law and Social Norms 66 (2000).

to go to the opera but not to watch a charitable organization feed the poor. Reputation theory, on the other hand, would predict that if donating to a particular cause is sufficiently visible then stable ILN will foster the provision of collective goods over social waste. In other words, the theory presented here unambiguously disqualifies burning money as a means of raising one's reputation – even if the fire was highly visible. Moreover, reputation theory predicts that the receivers of givings correspond to the welfare of the particular group, in which the respective donor aspires to reputation. Even though it is sometimes hard to tell whom a donator seeks to impress some cases are quite clear: It is not surprising that companies donate for their pressure groups or that law firms tend to support law schools (rather than, say, the protection of bio-diversity). The next subsection considers a case of selfish groups with a potentially devastating effect on society as a whole.

2. Selfish Groups: Racism

Under reputation theory, players contribute to collective goods because outsiders are relatively free to award or withhold reputation. Outsiders use this freedom to foster the collective goods that they appreciate but, without reputation norms, would not be supplied due to the collective action problem. It is clear from this analysis that if all relevant outsiders belong to a particular subset of the population the resulting norms will reflect their interest *as a group*, not that of society as a whole. Thus, if a certain type of transaction gives rise to a distinct type of reputation and if the transaction is confined to a particular subset of players, the respective group will harness its reputation norm to pursue its own selfish group

interest. It is not unlikely that such selfish groups impose externalities on other groups or on society as a whole.

Racism is a striking example for selfish group behavior. Again, it is instructive to compare the reputation theory account to that of Posner's signaling model. Posner sees discriminating on the basis of race as a costly act that can signal the racist's low discount rate. By contrast, McAdams has pointed to the fact that racist ideology often emphasizes the inferiority of the discriminated group. This, McAdams argues, is at odds with Posner's requirement that a signal appear costly (the more costly the better!); foregoing cooperation with inferior people does not cost much.[87]

Reputation theory has little difficulty in explaining racism, including the alleged superiority of the racist's own race. At variance with the signaling model, reputation theory predicts that it is easier to convince others of a racist norm if those discriminated against are *not* attractive cooperation partners. This is so because the less costly a reputation norm is the more easily it is implemented. As an example, take discrimination against African-Americans in the South of the United States. Given the relatively poor education of most African-Americans, whites did not forego large gains from cooperation when they declined to employ African-Americans in well-paid positions. On the other hand, racism did allow hiring African-Americans

[87] See Richard H. McAdams, Signaling Discount Rates: Law, Norms, and Economic Methodology, 110 Y. L. J. 625, 651-653 (2001). – To counter this objection, Posner argues that debasing the worth of individuals of the other race actually *emphasizes* the cost of discrimination by clarifying that discrimination is not a matter of indulging an individual taste. See Eric A. Posner, Law and Social Norms 139 (2000).

for physical work. Thus, the racism norm effectively created a white cartel for well-paid employment, which is a very obvious case of selfish group behavior.[88]

Yet reputation theory has to overcome a difficulty in explaining racism that signaling theory does not face. Reputation theory usually predicts norms to be of a general character.[89] Racism, by contrast, consists of discriminating against a subset of the population. Therefore, reputation theory predicts that racism must create a norm network that applies only to members of the racist's own race. That is, racism must create a separate group based on race distinctions and having a distinct reputation mechanism. Simultaneously, race boundaries must seem fixed lest other groups fear becoming the next target. Both conditions were easily met in the South of the United States as slavery had secluded African-Americans from the white population anyway. By contrast, the Nazis had to launch a long propaganda campaign to gain support for their racist assaults on German Jews. Although they could build on a long history of discrimination against Jews, their ultimate success in establishing their racist ideology turned on the fact that they obtained control of the government.

3. HIDDEN BENEFITS: FENCING DUELS IN GERMAN STUDENT FRATERNITIES

Racism conforms to reputation theory in that it benefits a selfish group by creating a cartel among its members. By contrast, a more serious challenge to reputation theory is ILN that seem not to benefit anybody. Reputation theory predicts

---

[88] See Andreas Moro & Peter Norman, A General Equilibrium Model of Statistical Discrimination, SSRN Working Paper 18 *et seq.* (2002).

that outside players in principle will use reputation to encourage supply of collective goods. This proposition is hard to reconcile with costly norms that seem not to produce any gain. To defend reputation theory for these cases requires showing that the norm, contrary to appearances, does produce a collective good for society or for the group holding the norm.

Fencing duels in German student fraternities provide an example.[90] The duels are an ancient tradition going back to the emergence of the fraternities in the first half of the 19th century. At variance with aristocratic dueling in Europe or the Southern States of the US, the fraternity duels mostly are not a sanction for an insult. Rather, they are conducted on a regular basis. Fighting a number of these duels is a mandatory requirement for student members of the fraternity. Students who decline to fight are expelled from the fraternity. While there are practically no lethal incidences in the duels, members frequently sustain injuries. Usually, these injuries are of a minor character but they sometimes cause lasting and visible scars. The rules of the duels deliberately exacerbate this risk by disallowing protection gear. Student members sometimes undertake to abolish or to evade the duels. Such attempts usually fail because senior members – who provide the funding for the fraternity – oppose the change.

---

[89] See *supra* at 21.

[90] In 1953 the German Federal Supreme Court (*Bundesgerichtshof*) ruled the duels to be permissible, ending a long-lasting debate. Even in this decision, the court suspected the duels of maintaining the customs and perhaps privileges of the upper class. See 1953 *Neue Juristische Wochenschrift* 473.

It is hard to see any benefit from the fencing duels for the members of the fraternities. On the contrary, the risk that members suffer lasting injuries seems to be a mere cost to the community, not to speak of the individual member. This suggests turning to the signaling model, which does not postulate that a norm benefits the group holding it. In the signaling framework, dueling may figure as a present cost on student members. Imposing such a cost may help to sort out members with a high discount rate who are too impatient to bear it. There is a problem though with this analysis. While being hurt in a duel inflicts present cost in the form of pain and other inconveniences it also imposes disutility in the future through the visible scar.[91] This cost structure significantly departs from signaling theory because signaling discount rates requires discriminating as much as possible between present and future cost.[92] The deviation from the ideal signaling cost structure does not come about accidentally. Rather, the fraternities that hold on to the tradition expressly deny adequate gear that would protect against disfigure ment.[93]

Since signaling seems not to provide a fully satisfactory answer, it is worthwhile to reconsider potential benefits for the group that would make it

---

[91] Having a scar in one's face used to be a clear indication of fraternity membership, which, at least in former times, was an advantage. This should not, however, lead one to conclude that scars were actually a (direct) source of utility. Rather, a norms theory must explain why a disfigurement that causes displeasure nonetheless confers a social advantage.

[92] It could be argued that present cost predominates in the form of fear. However, there are many other ways of inflicting fear without risking long-term harm.

[93] It seems to me that there is a number of norm-induced activities that put at risk future utility. For instance, while drug consumption is often thought of as lack of self-discipline there is also the story that drug consumption is a signal for courage and independence.

consistent with a reputation account of the dueling norm. In fact, the dueling norm may have a hidden benefit for the members of the fraternity. These benefits can only be seen by considering the overall economic structure of the fraternity. One important part of the deal is generous support to student members, especially in the form of cheap housing and subsidized leisure activities. Since the fraternity is funded by its senior members it must be cautious to deter free riders who reap the benefits without being prepared to provide financial support after graduation. This free rider problem offers a simple explanation for diverging interests between student and senior members with respect to the duels. However, the free rider problem does not explain the duels. As such, an "entry fee" may keep students from acceding to the fraternity but it can hardly prevent them from behaving opportunistically once they have paid the fee and reaped the benefits.

It is at this point that the second good provided by the fraternity comes into play. The fraternities do not confine themselves to supporting its student members. In addition to that, they are a network of trust for advice, assistance, and friendship. Fraternity members support each other in various ways after graduation. They form a network of professional as well as private contacts. Such cooperation rests on a high level of trust. Another way to put this is that the fraternity creates its own peculiar type of reputation. Such a peculiar reputation, or higher level of trust, must be safeguarded against hit-and-run strategies. Therefore, membership in the fraternity must not be cheap so as to discourage opportunistic behavior beforehand. The reason why fencing duels are good for the fraternity, thus, is the following: By extracting an "entry fee", the fencing duels force new members to invest into the fraternity. Later,

those members will live up to their obligations within the fraternity because this is the only way not to lose this investment.[94]

In sum, reputation theory appears twice in the analysis: On a first level, reputation theory explains that the duels are maintained (in spite of the fact that some members would benefit from abolishing them). The explanation, however, rests on the condition that there is a benefit for the fraternity as a whole. The second level of analysis finds this benefit in the cooperative trust within the fraternity. The maintenance of a strong cooperative reputation norm is a collective good that must be preserved by imposing the duels as an "entry fee" on new members.

V. Conclusions

Communication and information impact on people's norms and values. This is not only a republican ideal. It is a reality experienced by all kinds of political and religious groups, parents, politicians, and other "norm entrepreneurs." To account for the prominent role of communication, in my view, is one of the key advantages to be expected of reputation theory. I cannot see how a behavioral theory of norms was able to explain the substantial role of communication: It is quite plausible that people

[94] The idea is that people can acquire a reputation in the first place by making a sacrifice that seems unrelated to the type of cooperation that reputation is meant to facilitate. The same logic applies where reputation is sold, for instance in the form of goodwill associated with a firm or brand. See David M. Kreps, Corporate Culture and Economic Theory, in Alt & Shepsle, ed, Perspectives on Positive Political Economy 90, 108-11 (1990), Steven Tadelis, What's in a Name? Reputation as a Tradeable Asset, 89 Am. Econ. Rev. 548 (1999), and Steven Tadelis, The Market for Reputations as an Incentive Mechanism, 110 J. Pol. Ec. 854 (2002).

obey norms from force of habit or internalization; but there is little in the way of predicting which norms will come to be internalized or accustomed to.[95] By contrast, if norms are here because they lift rational indeterminacy, it is natural that communication – the act of producing common knowledge[96] – is of great relevance. The ways in which people shape norms in public and private discourse is a deserving subject for further study.

Beside positive and instrumental analysis, there is also a normative gist to norms theorizing. It has been inferred, from the existence of norms, that individuals have only rather limited capability for autonomous self-determination.[97] Thus, it has been argued that in the virtually all-encompassing realm of norms, preference autonomy cannot be used as a baseline for normative analysis. To refute, at least weaken, this argument, it has to be shown that norms are well consistent with rational choice – rational choice being at the heart of preference autonomy. This, in my view, is also an important point to be made in favor of reputation theory (as well as for all other rational choice accounts of social norms).

On the other hand, the normative implications may not be that far-reaching.[98] Although a rational choice explanation of norms helps defend anti-paternalism, it

---

[95] E. Posner raises a related point when he says that internalization lacks a well-developed theory of how and when feelings of guilt occur, see Eric A. Posner, Law and Social Norms 43 (2000).

[96] See n. 19 *supra*.

[97] See the references in n. 2 *supra* at 1 .

[98] See the comparison between the two camps in Robert C. Ellickson, The Market for Social Norms, 3 Am. L. & Econ. Rev. 1, 30 *et seq.* (2001).

cannot dismiss the possibility of inefficient norms. True, reputation theory does offer some reason for being optimistic in this regard: If people can speak out about their preferences, reputation norms tend towards enhancing efficiency.[99] However, nothing guarantees that the players who can influence the norm comprise the whole population. If they constitute only a subset of the population, the group's interest need not coincide with the welfare of society as a whole, or with that of other groups. Conflicts between groups can be especially grave when it comes to issues of distribution. In such cases, supporting one's own group can earn one reputation. We have seen that racism is a natural application of this theory.[100] It follows that, at this level of generality, the theory cannot determine if norms will lead to efficient results.[101] Thus, it seems that saving rational choice as a foundation of norms analysis would buy us only little: Instead of arguing that people do not possess genuine preferences at all, advocates of strong government can invoke the (allegedly) detrimental effect of some inefficient norm. Also, government intervention against bad norms may be justified even against the protest of those individuals who suffer from the norm: A norm may be so entrenched that even its victims are forced to oppose change. One example is gender roles in certain cultural environments where women are barred from education. In spite of their inferior role, many women assert that they are content with their cultural background, including their role. Even if

[99] At least, they tend towards eliminating inferior norms. It is a different matter if there is sufficient supply for efficient norms.

[100] See *supra* IV.2 at 52

paternalism cannot be defended theoretically, we may have, in such situations, strong grounds for government intervention *as if* paternalism was legitimate.

Thus, even under a rational choice theory of social norms it may be hard to establish if an individual's statement reflects her preferences, or if it is forced on her by a norm. The goal to distinguish the two possibilities leads us to favor certain democratic rules of procedure, which shield the influence of detrimental norms to the best extent possible. To provide procedural rules like secret voting is, therefore, one normative implication of rational choice norms analysis. Where such rules are infeasible, rational choice and preference autonomy at least instructs the analyst to conduct a normative gedankenexperiment: If we subtracted the particular norms affecting an individual, what norms would she choose to live under? To pursue this question both defines and advises a research agenda for normative "law and norms:" That norms are the result of communication and rational choice, and nothing more, tells us what to substract, and hence what to look for.

---

[101] The same skeptical conclusion is drawn by Eric A. Posner, Law and Social Norms 176-7 (2000). See also Richard A. Posner & Eric Rasmusen, Creating and Enforcing Norms, with Special Reference to Sanctions, 19 Intl. L. & Econ. Rev. 369, 380 (1999)