Synthese (2013) 190:585–618
DOI 10.1007/s11229-012-0194-8

A convention or (tacit) agreement betwixt us: on reliance and its normative consequences

Luca Tummolini · Giulia Andrighetto · Cristiano Castelfranchi · Rosaria Conte

Received: 15 August 2012 / Accepted: 15 August 2012 / Published online: 29 September 2012
© Springer Science+Business Media Dordrecht 2012

Abstract The aim of this paper is to clarify what kind of normativity characterizes a convention. First, we argue that conventions have normative consequences because they always involve a form of trust and reliance. We contend that it is by reference to a moral principle impinging on these aspects (i.e. the principle of Reliability) that interpersonal obligations and rights originate from conventional regularities. Second, we argue that the system of mutual expectations presupposed by conventions is a source of agreements. Agreements stemming from conventions are "tacit" in the sense that they are implicated by what agents do (or forbear from doing) and without that any communication between them is necessary. To justify this conclusion, we assume that: (1) there is a salient interpretation, in some contexts, of everyone's silence as confirmatory of the others' expectations (an epistemic assumption), and (2) the participating agents share a value of not being motivated by hostile attitudes (a motivational assumption). By clarifying the relation between conventions and agreements, the peculiar normativity of conventions is analyzed.

Keywords David Lewis · Convention · Norm · Tacit Agreement · Confirmatory bias · Pragmatics

A convention, like the regularity of 'driving on the right' in Italy, is a social means for the sake of a common end or interest. A common end need not be a goal or desire

[1] Two agents "agree in desires if exactly the same world would satisfy the desires of both; and a world that satisfies someone's desires is one wherein he has all the properties that he desires de se and wherein all the propositions hold the he desires de dicto. Agreement in desire makes for harmony." Lewis (1989, p. 119). On the distinction between attitudes de dicto and attitudes de se see Lewis (1979a).

L. Tummolini (✉) · G. Andrighetto · C. Castelfranchi · R. Conte
Istituto di Scienze e Tecnologie della Cognizione (CNR), Via San Martino della Battaglia 44, 00185 Rome, Italy
e-mail: luca.tummolini@istc.cnr.it

that we pursue together intentionally (e.g. a shared intention to drive on the right). A set of goals that are co-realizable may suffice (e.g. our self-regarding goals to avoid collisions in traffic): a common end is, at least, an *agreement in desires*.[1] Describing a convention is describing a way in which we regularly behave in a certain class of situations, which is sufficient to obtain something we all want. Thus, a convention is a form of joint action but without a shared intention to act in that way. However, conventions are not necessary means. They are arbitrary regularities since some other way of behaving might serve the same common interest, e.g. we might as well jointly drive on the left to avoid colliding. That is, our common interest (our ends in agreement) can be fulfilled if our choices of the means are also in agreement, when at least another possible arrangement of choices is foreseeable. To be useful, conventions should be stable: once established, conventions perpetuate themselves. And they do so because it is in the best interests of all of us to keep acting as we do, if others do the same. Moreover this, along with everything else described here, is common knowledge between us, so much so that if one bothered to engage in explicit reasoning from the perspective of another fellow, one would discover that conformity to the convention is in the best interests of everyone else as well as of oneself, and so be assured that the regularity will persist.

The idea that conventions along these lines are a peculiar kind of regularity in the behaviour of a population of agents has been forcefully defended by Lewis (1969; see Sect. 1 below), who considers his theory analogous to the one sketched by Hume (1740) in the *Treatise of Human Nature* while discussing the origin of justice and property.

According to this view, conventions *describe* a self-enforcing behavioural pattern; do they *prescribe* it too?

Many critics of Lewis's theory of conventions have pointed out that Lewis seems to miss the *normative component* that is a part of conventions. One way to express the critique is to say that *conventions are not mere regularities but rules*, not only regularities de facto but also regularities de jure (Postema 1982; Gilbert 1989; Marmor 1996). That is, once a conventional regularity to drive on the right is established, driving on the right is not only what we *usually* do, it is something we *ought* to do. And the same is true for all the conventions we are parties to. Conformity to our conventions is not just what we happen to do, it is something that is *required* of us.

One natural way to explain this normativity of conventions is to view conventional regularities as a form of *agreement* between the agents recurrently acting together (Gilbert 2008). Besides, Hume himself has suggested that a convention is an "agreement betwixt us, though without the interposition of a promise" (1740, Book 3, Part 2, Sect. 2). But if a convention is often just a collectively *unintended* kind of behaviour while agreements seem to be a joint *intentional* action *par excellence*, how can they be related? And, more importantly, why does an agreement have normative consequences in the first place?

However, though often unacknowledged, Lewis has indeed offered an account of the normative dimension of conventions, and in a way that is, at first sight, different from an appeal to agreements. Lewis argues that: "any convention is, by definition, a *norm* which there is some presumption that one *ought* to conform to (…) it is also by definition a *socially enforced* norm: one is expected to conform, and failure to con-

form tends to evoke unfavourable responses from others" (Lewis 1969, p. 99; emphasis added).

What kind of norm any convention is, however, is not immediately clear.

Lewis suggests that there may be all sorts of reasons why, for a *particular* convention, one 'ought' to conform to that particular regularity. If the convention originated in an exchange of promises, then one also 'ought' to act to keep the promise; if the convention is also a social contract, then one 'ought' to reciprocate the obtained benefit. But, Lewis contends, there are also *general reasons why one ought to conform which are valid for any regularity that qualifies as a convention*, for any population relative to which the convention exists, and for any situation the convention applies to.

Such general reasons derive from the fact that, by conforming to a convention, (1) one acts in one's own best interest, and, at the same time, (2) in a way that answers to others' preferences, *when they reasonably expect one to do so*. Both *acting in one's own best interests* and *in a way that is in the interest of others (when they reasonably expect one to do so)* are something that, according to Lewis, "we do presume, other things being equal, that one ought to do" (Lewis 1969, p. 98). If the former is a requirement of instrumental rationality, the latter stems from a *moral principle* that we, somehow, accept.[2]

But is it so?

Alice expects Bob to carry out an action because Chris, who is honest and reliable, has told her that Bob intends to do it. She completely trusts Chris; hence Alice has good reason to believe what Chris says. Moreover, she really wants Bob to act in that way: Bob acting according to her reasonable expectation answers her preferences. Is this sufficient for Bob to be 'required' to carry out the action in question? If Bob has not in any way induced what Alice believes, why 'ought' he to carry out that action?

Similarly, one can be reasonable in expecting conformity to a certain convention given widespread conformity in the population. For instance, it is reasonable in Italy to expect the driver of the oncoming car to keep to the right, given one's experience with what Italian drivers usually do, even without any direct experience with the particular others one is now dealing with (i.e. one's expectation about what the next driver will do is not grounded in one's own experience with that driver). How is it, then, that such an anonymous agent is responsible for expectations he has not himself induced? Though it seems that, notwithstanding what an anonymous driver will in fact do, any driver 'ought' to conform to the convention that prevails in that population; it is not evident why any fellow is so bound, at least *when he bears no direct responsibility for what anyone reasonably expects from him*.

In order to clarify what kind of normativity characterizes any convention we aim to show in this paper that the moral principle strategy and the agreement strategy can be fruitfully combined. First, we argue that conventions have normative consequences in relation to the role that trust and reliance play in their maintenance. We contend that it is by reference to a moral principle impinging on these aspects (i.e. the principle of Reliability) that interpersonal obligations and rights originate from conventional regularities (see Sects. 3 and 5). Second, we argue that conventions are sources of

[2] See Sugden (2000, 2004) for a development of Lewis's approach to the normativity of conventions, and Verbeek (2002) for a critique of Sugden's theory.

agreements, even if it is not necessarily by agreement that a convention is established. It is usually argued that once the notion of convention is understood, it is thereby clarified in what sense a behavioural regularity is also an agreement. But agreements are not only agreements in desire that, as a consequence, produce regularities in behaviour. In fact, an agreement is also a specific type of social relation between agents that is intentionally created to produce such agreements in desires (see Sects. 6.2 and 7). Agreements are considered by Lewis to be one of the possible means of producing a system of mutual expectations (1969, p. 35), but the converse also holds, or so we argue: a system of mutual expectations of the kind presupposed by a convention is a source of agreements (Sect. 8). This suggestion is at first, however, counter-intuitive given that conventions are typically maintained without the need for any communication between the parties. If this is true, how can agreements be established without communication? How can conventions be sources of real agreements and not a way of behaving as if we had agreed even though we hadn't?

By clarifying the relation between conventions and agreements, the peculiar normativity of conventions is also analyzed. We argue that the normativity of conventions is the same normativity of agreements, and that conventions are the sources of agreements, *tacit* agreements but agreements nonetheless.

1 Convention: a definition

In order to explore the normative dimension of conventions, it is useful to start by offering an explicit account of what a convention is.

A few years after his first contribution on the topic (Lewis 1969), Lewis amended his original analysis by offering the following definition, which we adopt in what follows (Lewis 1975, pp. 164–165):

A regularity R, in *action*, or in *action* and *belief*, is a *convention* in a population P if and only if, within P, the following six conditions hold:

1. Everyone conforms to R.
2. Everyone believes that the others conform to R.
3. This belief that others conform to R gives everyone a good and decisive reason to conform to R themselves.
4. Everyone who believes that at least almost everyone conforms to R will want the others, as well as himself, to conform.
5. R is not the only possible regularity meeting the last two conditions. There is at least one alternative R^i such that the belief that others conformed to R^i would give everyone a good and decisive reason to conform to R^i likewise.
6. Finally, the various facts listed in conditions (1)–(5) are matters of common knowledge.

This definition is meant to capture the core of our common concept of convention whereby we are ready to acknowledge that *a practiced regularity* of action (or of action and belief) (condition 1), that *everybody expects widespread conformity* to (condition 2), that is *serving some common end of ours* (condition 4), that is *arbitrary* (condition 5) and *self-enforcing* because it is openly known (condition 6) that expected

conformity gives everyone a reason to go on conforming (condition 3), is what we would indeed consider one of our conventions.[3]

Lewis amended his more famous game theoretical definition in several ways, but one change is particularly relevant to his original target: to understand in what sense language is a convention-governed human activity. Since clause (3) was originally formulated in terms of a 'conditional preference' for conformity, the only regularities that were considered conventions were regularities in action alone. In other words, given that one cannot *conditionally prefer to believe* something because one cannot *deliberatively choose* what to believe, regularities in action and belief were excluded. As a consequence the convention governing the use of a language was characterized as a convention of *truthfulness* in a given language, and the agents conforming to the convention of language are its speakers alone. In this way, present speakers are viewed as coordinating with speakers who truthfully used that language in the past (Lewis 1969, p. 194).

However, adopting the second definition, *regularities in action and belief* can also be considered conventions. From this perspective, others' expected conformity provides one with a *reason* either *to do* (a practical reason) or *to believe* something (an epistemic reason). Thanks to the formulation in terms of reasons for conformity (instead of preferences for conformity), the convention governing a language coordinates speakers with listeners, and not speakers alone. More specifically, according to Lewis, the convention, whereby a population uses a language, is a *convention of truthfulness and trust*, that is, a regularity in which conformity amounts to speakers *doing* something (i.e. speaking truthfully) and to listeners *believing* what speakers say (i.e. trusting). Given that both speakers and listeners share an interest in communicating, each other expected conformity is a practical or an epistemic reason for their own conformity.[4]

2 Trust by convention: when trust is based on trust

Once, however, conventions are viewed as regularities characterized by these features, it is also clear that *trust*, properly understood, is not peculiar to conventions of language alone.

Trust, in fact, is not only the action of trusting another agent. It is also a complex state of mind in which an agent *expects* and *wants* another agent to do something, *relies on* this agent to behave in this way, and *delegates* the fulfilment of the agent's goal to another agent (Castelfranchi and Falcone 2010).

[3] Over the years, many have challenged Lewis's analysis under several different aspects. Notwithstanding this, and especially adopting the 1975 formulation, we think it is still the most fruitful way conceptualizing social conventions. In this paper we focus on how, having fixed the concept of convention, to account for the normativity of convention and its relation to agreements. For an extended critical assessment of Lewis's theory see Gilbert (1989). For a recent critical re-appraisal of Lewis's theory see the contributions in Tummolini (2008).

[4] Lewis's specific analysis of conventions of language has been widely criticized often on the ground that truthfulness is not a convention in actual language use but a norm (see for instance Jones 1983).

Crucial for trusting is 'reliance *on an agent for something*', and not just reliance on something happening (Holton 1994). When we rely on something happening, say that the train will arrive on time, we assume that it will happen (usually because we believe that it will happen), and plan or intend accordingly. Differently, *when we trust an agent, we rely on him as a 'cognitive' agent*, that is, as an autonomous entity whose behaviour is caused by his own beliefs and goals. Hence, when *we rely on an agent* to behave in some way, *we plan or intend what to do on the basis of the belief that such an agent will behave in that way on the basis of his reasons* (her beliefs and goals), and not, for instance, because he is coerced into behaving that way. For example, if Bob coerces Alice into giving him her purse, Bob relies on the fact that Alice will give him the purse but he does not rely *on* Alice to give it to him; there's no question of trust in coercive interactions. Trust is a fundamental non-hostile attitude and behaviour.

Trust, moreover, is not without peril, since, by trusting another, the agent makes himself *vulnerable*: he exposes himself to the risk that the other will not behave in the expected way, thus frustrating his desires. Actually, trust also presupposes that the trustee too is not motivated by a hostile attitude towards the trustor, or at least that *the trustor expects that those he relies on are not hostile* (Castelfranchi and Falcone 2010).[5]

More generally, trust is always relative to a goal whose fulfilment *depends* on another agent's behaviour.[6] Goals can be either *epistemic* ones (i.e. the goal to know something or to know whether something is true or not) or *practical* ones (i.e. the goal that the world be in some way or to carry out an action). Correspondingly, reliance can be either for an epistemic or for a practical goal.

That is, if one *epistemically* relies on another agent, one relies on another agent in order to fulfil one's own epistemic goal, something that typically happens through communication. In such a case, epistemic reliance entails that the trustor believes that the trustee will truthfully communicate with him because the trustee is motivated (for some reason) to act in this way. Epistemic reliance on another agent entails coming to believe what the trustee wants the trustor to believe. This is the kind of epistemic trust that Lewis had in mind by introducing the convention to trust the speaker.

On the other hand, when the trustor relies on the trustee *practically*, the former relies on the latter for fulfilling one of her desired states of affairs. If Alice relies *on* Bob to drive on the right side of the road, Alice believes that Bob will drive in that way because Bob has a reason to carry out such an action and Bob's behaviour fulfils her goal.

In both situations, by coming to believe something or by acting on the basis of one's expectation about another agent, the former trusts the latter.

Finally, one trusts on the basis of reasons. Sometimes trusting may be 'irrational', such as when by making oneself vulnerable, one thereby creates a selfish reason for another agent to exploit that vulnerability.[7] At other times, trust is perfectly reason-

[5] See Sect. 3.1 for the relevance of not being motivated by hostile attitudes.

[6] That is, the trusting agent believes herself to be *dependent* on another for the fulfilment of one of his goals, see Conte and Castelfranchi (1995) and Castelfranchi and Falcone (2010).

[7] If one extends a loan to another assuming that the other will do his best to repay it, one also gives the other a selfish reason not to repay it; see Bacharach and Gambetta (2001).

able, such as when one relies on another to do something simply because it is also in the interest of the other agent to act in that way. Even if in this latter case trust is reasonable and safe, it is not of course without risks given that the other could simply *change his mind* and act differently.

What is, at this point, the relation between trust and conventions?

According to the definition of convention given above, in any convention, the agents want the others to conform, expect future conformity from their fellows, and this belief is a reason for everyone to conform (i.e. a practical reason to carry out an action or an epistemic reason to believe something). Given this, it is also clear that *any act of conformity to a convention presupposes reliance on the others to conform*: the reason for conformity is also the reason to trust, i.e. to rely on others and to do something accordingly. Since, however, by conforming one trusts in others' conformity, that is, in their trust in oneself, *conventional regularities are regularities of reciprocal trust*. Moreover, in conventions, *trust is based on trust* because the expectation of conformity, grounded in past conformity, is a reason to conform in the present: Alice has a reason to trust Bob if Bob trusts Alice, and likewise for Bob.

More precisely: a regularity R in *reciprocal trust* in a population P is a *convention* if and only if the following six conditions hold:

1. Everyone regularly trusts each other while conforming to R.
2. Everyone believes that the others trust each other as well as oneself when conforming to R.
3. This belief that everyone trusts each other in conforming to R gives everyone a good and decisive reason to trust the others while conforming to R themselves.
4. Everyone who believes that the others conform to R (i.e. reciprocally trust each other) will want the others, as well as themselves, to conform (i.e. to trust oneself).
5. R is not the only regularity meeting the last two conditions. There is at least one alternative regularity R^i in reciprocal trust, which would perpetuate itself instead of R.
6. The various facts listed above in conditions (1)–(5) are matters of common knowledge.

Any convention, then, is always a form of reciprocal trust, which is sustained by past reciprocal trust, and that breeds future trust. In particular, the belief that everyone reciprocally relies on each other is a reason for everyone to rely on the others. This reason can be for *practical reliance*, if conforming to R is a matter of reliance on the others to act in a certain way, and acting oneself accordingly. The reason can be for *epistemic reliance*, if conforming to R is a matter of reliance on the others to act in a certain way, and believing oneself accordingly. In the case of *a regularity of practical trust*, some desired end might be reached by relying on others, provided that the others also rely likewise; therefore one wants to rely on others, if they rely on one. In the case of *a regularity* in which *epistemic trust* is involved, one's belief, together with the belief that the others practically rely upon one, are premises that deductively imply or inductively support a conclusion, and by believing this conclusion one would thereby conform to R (i.e. he would epistemically rely on the others). Such reciprocal trust is reasonable since by trusting each other we are able to agree on the choice of the means

to fulfil each of our individual ends. This agreement, however, is only agreement in desires.

Reciprocal trust can originate in several different ways, for example by explicit agreement. However, a regularity of reciprocal trust qualifies as convention in the way it perpetuates itself, and not in the way it originates. Hence, there is *trust by convention whenever it is our reciprocal trust that, together with our desires for the end, gives us a reason to keep on trusting*.

Generalizing the definition of convention in this way is faithful to Lewis's analysis because no modification or additional clause has been proposed. Whether highlighting the role of trust and reliance is also fruitful in understanding the peculiar normativity of conventions is explored in what follows.

3 The normative consequence of reliance: the principle of Reliability

It is important to notice, at this point, that there seems to be a general reason why inducing trust and reliance on oneself and then disappointing it is doing something *wrong*.

One way to clarify why this is so is by appealing to those moral principles that focus on "what we owe to each other when we have led them to form expectations about our future conduct" (Scanlon 1990, p. 200).

With the aim of explaining the normative consequences of promises, Thomas Scanlon, in his seminal contribution, has proposed several moral principles that all concern the elicitation of expectations in others.[8] Consider, for instance, the so-called *principle of Loss Prevention* (Scanlon 1990, p. 204). This principle requires that "one that has intentionally or negligently led someone to expect that one will follow a certain course of action, and has reason to believe that person will suffer significant loss as a result of this expectation if one does not fulfil it, must take reasonable steps to prevent that loss, that is, he *ought* to warn, fulfil the expectation or compensate" (emphasis added).

Though encouraging expectations in others is not the same as inducing their reliance, since the principle of Loss Prevention is not just aimed to prevent another agent's desires frustration but his *losses*, some form of reliance seems to be presupposed for the principle to be applicable. For instance, suppose that Alice needs Bob's car tomorrow, and that Bob has led her to expect that he won't need it. Bob is aware of Alice's expectations because he knows that she heard him accepting a lift from a colleague on the phone. Suppose, however, that Alice decides not to rely on Bob's decision, and that she books a taxi to be completely safe. Knowing this, Bob is under no obligation towards Alice, not even to warn her if, eventually, he decides to take the car. Though taking Bob's car might be something she prefers over taking a taxi, the frustration of her desire to go by car is not a loss Alice incurs: it is not something that she has and wants, which she is deprived of. Hence, the principle of Loss Prevention does not apply. In this respect, even if she had relied on Bob, the frustration of her desire if Bob

[8] Scanlon's aim is to offer an account of the wrong of breaking a promise (seen as a device for eliciting expectations) by introducing and justifying a moral principle of Fidelity (1990, p. 208). However, this principle may be unnecessarily strong when trust and reliance are not based on promises, which is the focus of this work.

takes the car is not a real loss.[9] However in relying on Bob, Alice has lost something she had: she has paid *opportunity costs* (i.e. the available alternatives of actions she had, and which she has missed by relying on Bob's car being available).

At least for the aim of this work, then, the way Judith Thomson has defended a similar principle seems better suited (Thomson 1990, p. 302). Thomson argues for the validity of a *Word-Giving Thesis* in which, when *an agent invites another one to rely on the truth of a certain proposition, which invitation the latter agent accepts* (or uptakes), then the latter agent acquires a *claim* (i.e. a right) *against* the former one to its being true. This way of formulating the moral principle bears three main advantages over Scanlon's: (1) it makes the relevance of *reliance* or *uptake* in this process explicit, (2) it generalizes the principle towards whatever proposition one may rely on besides those that refer to an action that one will do in the future, and (3) it clarifies that the involved obligations and rights are *directed* (i.e. the normativity characterizes a relation between a bearer and a counterparty whom the bearer is bounded with).[10]

However, though one can *induce* reliance, one can *allow* reliance as well, and in such a way that has normative consequences.

Consider again the previous example: Alice has heard Bob's conversation with somebody else and, as a consequence, she comes to believe that Bob will not take the car, and she therefore relies on it. In this case, Bob has unintentionally induced in Alice some kind of reliance. We have suggested that by acting on these expectations about Bob, she will incur some losses, and so the principle of Loss Prevention might apply. But is it so? After all, such induced reliance in this case is not intentional; can Bob be responsible for Alice's unilateral decision to rely upon him in this situation?

It seems correct to say that though her reliance has been unintentionally induced, Bob has at least 'allowed' her to rely on him. More precisely, *allowing* a belief or an action is *to have the power to disconfirm another's belief* (which is a reason to believe something else or to act in some way) *and forbearing to disconfirm it*. If hearing what Bob has said on the phone is a reason for Alice to believe that he will not take the car tomorrow, then this belief is obviously something that Bob can disconfirm. By not disconfirming such belief, Bob is also allowing her to believe in this way.

Granted this, this form of allowing as such is still not sufficient for an agent to acquire a right against another. Suppose that, immediately after having realized her reliance, Bob clarifies that what has happened just means that he has not confirmed that he will take the car (which is the same as not disconfirming the belief that Bob will not take it) and nothing more than that. Can Alice hold him responsible for her losses if, in the end, he decides to take the car despite her unilateral reliance? It seems not.

But suppose that after his conversation on the phone, and knowing that she needs the car, Bob turns to Alice and say 'yes, you heard correctly. I won't take it!'. By

[9] Though it can be so when one considers the desire to have the car not as something one will achieve but as something already achieved and to be protected; see for this possibility and its psychological plausibility (Miceli and Castelfranchi 2002).

[10] Gilbert (2004) has refuted Scanlon's account of promissory obligations precisely for failing on this last dimension. For the notion of *directed* obligations, obligations in the relational context of a bearer and a counterparty see Hohfeld (1996); for a formal model of this kind of obligation see Henning and Krogh (1995).

confirming a belief that he has unintentionally induced in her, Bob then becomes *obliged towards* her, to warn her if he changed his mind, or, if it's too late, to do as expected or to compensate. Because such confirmation of the belief logically entails the absence of a disconfirmation, even in this case Bob has allowed her to believe something, though not 'passively' (i.e. by forbearing to disconfirm it) but 'actively' (i.e. by confirming it). It is this form of *active allowing* that is sufficient for the moral principle to apply when one does not intentionally induce reliance in others.[11]

Finally, there are also cases in which one actively allows other agents' reliance on oneself in ways that one has not induced.

Suppose for example that Alice believes that Bob will not take the car tomorrow because Chris told her so, and that she relies on him for use of the car. Bob knows about all this, and he allows her to believe it by forbearing to disconfirm such belief. If she just acts on this basis (not knowing that Bob knows about her reliance), it seems that at most Bob should warn her if the belief is false, but if this is so, it is just out of sheer altruism.[12] If, on the other hand, Bob has confirmed her expectation, for instance by nodding, Bob has actively allowed her to rely on him, and, from there on, he is responsible for her possible losses even if he has not directly induced that belief in the first place. Again, when it's too late for a warning, Bob *ought* to fulfil the expectation or compensate.

Hence, when an agent intentionally induces or actively allows another agent's reliance on oneself, the former agent undertakes *a duty of reliability* towards the latter agent, and creates a corresponding *right to rely*. Reliability is normatively required to prevent losses caused by intentionally inducing or actively allowing such reliance. One way in which such a principle can be explicitly formulated is the following: *if one intentionally induces or actively allows another agent to rely on the truth of a certain proposition, then the latter agent acquires a right to reliability (i.e. to be warned if the proposition turns out to be false, or, if the proposition is about the future action of the former agent and it is too late for a warning, a right that the former agent acts so as to make the proposition true or compensates for the incurred losses)*. For these reasons, we name such a principle: *the principle of Reliability*.

3.1 The principle of Reliability and the value of non-hostility

So far, we have suggested that if the agents endorse the principle of Reliability they are disposed to acknowledge that specific obligations and rights arise as a consequence of intentionally inducing or actively allowing reliance from an agent. Exploring, however, under what conditions the agents endorse this principle is a different question.

To understand these conditions, we should notice, first of all, that if one is disposed to merely ignore the losses caused to others, *when one is responsible for such reliance*

[11] Scanlon's principle of Loss Prevention indeed also mentions leading expectations 'negligently', besides doing it intentionally. However, negligence implies having not paid *due* care to avoid such reliance, and so it cannot be evoked to explain, without circularity, a principle which normatively demands such behaviour. On the other hand, the notion of active allowing has no such problem.

[12] The reason why common knowledge of an agent's silence (i.e. the forbearance to disconfirm another agent's beliefs) may change the situation is discussed in Sect. 4.

(i.e. one has intentionally induced or actively confirmed the reliance on oneself), either one desires that others do in fact incur those losses, and in so doing *one is disposed to act with hostility towards the others*, or at least one lacks the desire that others do not incur them, that is, *one lacks a disposition to act in a non-hostile way*.

Thus, assuming that *the agents share a value of not being hostile in their social interactions*, and especially in their *joint actions*, would also entail that agents endorse the principle of Reliability, because such a principle can simply be seen as an instantiation of the general value with respect to the special case of reliance and trust.[13]

Moreover, at least according to Lewis' dispositional theory of values (Lewis 1989), the value of non-hostility is a *value* de se, that is, *a property that the agents of a relevant population are disposed to desire to desire (i.e. to value) under ideal conditions*.[14] Adopting this view, *the value of not being hostile in their joint actions* amounts to the fact that, *if the agents of a relevant population are under ideal conditions, they are disposed to desire to desire to have the property of not being hostile*: i.e. they desire to desire to refrain from acting in a hostile manner. As a consequence, given that being hostile is being motivated to frustrate the desires of another agent, complying with such a value also entails that one is disposed to revise one's possible first-order hostile desires in a way that, if done by all, would inevitably result in the creation of harmony in the population, that is, in *agreement in desires*.[15]

It is important to notice, moreover, that sharing this value does not necessarily imply that the agents *will* behave in the present conditions in the way they *would* if they *were* in ideal ones. However, in order to engage in a social interaction, the agents must at least share an *expectation* that each of them *is* so motivated, otherwise the best that they can do is to avoid any possible contact with each other.

Finally, viewing values with the help of Lewis' dispositional theory has the important advantage of also offering a way of *naturalizing* them, since the fact that a property or a state of affairs is a value is explained in terms of dispositions and mental attitudes: the disposition to have certain second-order motivational states under ideal conditions. A dispositional theory of value, in Lewis's words, "reduces facts about values to facts about our psychology" (1989, p. 113).

[13] The principle of Loss Prevention is justified by Scanlon within his contractualist approach by the fact that "it is not unreasonable to refuse to grant others the freedom to ignore the losses caused by the expectation they intentionally or negligently lead others to form" (1990, p. 204), and the agents in a population have a reason to refuse such freedom when they share the value of assurance (p. 206). Scanlon defines 'assurance' as "being able to be reasonably certain that a thing will happen unless one consents to its not happening" (1990, p. 222). For a more general discussion of this brand of contractualism see Scanlon (1982, 1998).

[14] Ideal conditions are those that one can put oneself in by "the fullest possible imaginative acquaintance that is humanly possible" and "the canonical way to find out whether something is a value requires a difficult imaginative exercise" (Lewis 1989, pp. 121–122).

[15] Hence, if the value of non-hostility is analyzed dispositionally, Scanlon's contractualism is not needed because failing to conform to the principle of Reliability would render one's behaviour hostile, and this is contrary to what the agents value.

3.2 On refraining from acting in a hostile way as a social contract

Though some level of presumptive non-hostility seems required at least to ease in-group interaction,[16] human societies most probably vary in the level in which their members value non-hostile interactions. But even a society in which this value is spread may suddenly abandon it, as when an unexpected catastrophe leads to social disintegration and diffuse reciprocal hostility. Thus, assuming that the agents do already share such a value leaves open the question of *whether there is any reason to endorse it in the first place.*

To see where this reason comes from, let's confine ourselves to the case of social interactions that are an instance of joint actions. In such cases, desiring not to be motivated by hostile attitudes is reasonable only *conditionally*: one has a reason to refrain from acting on hostile motivations *if* the others are expected to do the same. That this is the case is especially clear when aiming to enter into a joint action, since being motivated to thwart the other agent's goals disrupts the effectiveness of the joint action itself. Hence, in order to have a reason to refrain from hostility when taking part in a joint action, one needs to *expect* the same attitude from others.

From this angle, then, the agents have a reason to value non-hostility when, in the population, a *regularity* prevails in which (1) all the agents refrain from being hostile when acting jointly, (2) everyone expects such regularity in refraining from hostility, and (3) such belief gives everyone a practical reason to refrain from acting with hostility towards others. Moreover, (4) everyone who believes that at least almost everyone regularly refrains from acting with hostility will want the others, as well as himself, to refrain from hostility, and (5) all these conditions are common knowledge between the agents. Under such conditions, a regularity acquires *stability* in that population.

It is not, however, a conventional regularity (strictly speaking) because, even if an alternative behavioural regularity of reciprocal hostility is indeed possible (and its awareness gives everyone a reason to be hostile), *the agents do not desire that the others conform to this regularity in reciprocal hostility*. In other words, the regularity in refraining from acting with hostility lacks the distinctive *arbitrariness* of conventions.

But, if one views the stable state in which everyone conforms to the regularity of non-hostility in their social interaction (and especially when acting jointly) as the *status quo*, and the alternative stable state of regular presumptive hostility as *the state of nature*, then *this regularity* (even if it fails to be a convention) *is an example of social contract* as proposed by Hobbes, Locke and Rousseau.[17] It may be a weaker social

[16] See for instance Hare (2007) who explores a similar behavioural trait and its role in the evolution of human cooperation. A certain level of "tolerance" is seen as prerequisite to enable the agents to reap the benefits of mutually advantageous joint action.

[17] Contrasting the concept of 'convention' with that of 'social contract', Lewis defines the latter as "any regularity R in the behavior of members of a population P when they are agents in a situation S, such that it is true, and common knowledge in P, that (1) Any member of P who is involved in S acts in conformity to R. (2) Each member of P prefers the state of general conformity to R (by members of P in S) to a certain contextually definite state of general nonconformity to R, called the *state of nature* relative to social contract R" (Lewis 1969, p. 89).

contract than the core traditional examples[18] because no one actually gains from being presumptively hostile with others if they are indeed non-hostile with themselves. However, this social contract is no less fundamental in that it is the most general requirement needed to keep a population and its groups together.

If we assume that a regularity with these features characterizes the population, then, the agents have indeed a reason to value non-hostility, that is, a reason to desire not being motivated by hostile attitudes when acting jointly. On this basis, when a social contract such as this is indeed established within the population, the moral conclusions of the principle of Reliability would actually follow directly from the fact that the agents have reason to value their reciprocal non-hostility. Regularities like that of a social contract of reciprocal non-hostility is all we need to ground the principle of Reliability.[19]

4 The ambiguity of silence and tacit confirmation

Even without accepting the dispositional theory of values, having identified the peculiar content of the moral principle of Reliability and the specific condition in which the agents may acknowledge its authority (i.e. when they share a value of non-hostility) is nevertheless relevant to understand the normativity of conventions.

However, assuming that the agents, who conform to a prevailing conventional regularity, share the value of non-hostility is not by itself sufficient to explain the normative consequences of conventions. Though a convention can fruitfully be seen as a regularity in reciprocal reliance and trust (see Sect. 2), the most typical reliance is usually neither intentionally induced nor actively allowed because often the agents do not rely on the others on the basis of direct past experience with each other.

However, consider, for the moment, this simpler scenario. Suppose that it is common knowledge between Alice and Bob that Alice wants Bob's car tomorrow morning, and that she believes that he will not take it because tomorrow is Monday, and on Mondays Bob never takes it (maybe just because it is his habit to act in this way or because the traffic on Monday mornings is more intense than on other days and Bob hates to be stuck in traffic). Given that Alice believes that Bob has his own reasons for not taking the car, and that she knows that he usually acts in this way every Monday, it is reasonable for her to expect Bob to behave in this way this Monday too (i.e. she believes with some degree that this event will happen). Moreover, assume that Alice is confident enough in this event to rely on Bob not taking the car, and she decides to travel to a meeting with his car. All above being common knowledge between them, *she also observes that Bob has been silent about the truth of her belief* (i.e. he has not disconfirmed it) until Monday morning. However, just when the time comes, Bob decides to take the car, perhaps because that particular Monday he needs the car for some unanticipated errands. Has Bob done something *wrong* to Alice?

[18] The core examples of social contracts are situations in which an agent may further his own welfare by violating the social contract when the others conform to it, but not if all behave in the manner of the state of nature. The latter is the worst state for all.

[19] For a more general discussion of the relation between conventions and moral norms along similar lines see Verbeek (2008).

It is foreseeable that having incurred some losses, Alice may resent Bob's last-minute decision, and she may even protest about such a sudden change of mind. But, is she *entitled* to anything? Is Bob under any sort of *duty* towards her? If she did think like this, Bob could legitimately claim not to have intentionally induced her reliance on him, not even to have acted in order to make her believe anything about him. So why would be Bob responsible for Alice's losses?

Something strange has happened.

The *closer* Alice comes to the fulfilment of her expectation that Bob will not take the car, the *surer* she feels about such fulfilment and *entitled* towards the other acting as expected. It is a fact indeed that, though Bob knew about her belief, he has been *silent* until the moment has come: Bob has not disconfirmed her belief.

Suppose that Alice has interpreted Bob's silent behaviour as a *confirmation* of her belief that he won't take the car. What kind of confirmation is this, given that they have not communicated with each other? Is it reasonable to read the other's silent behaviour in this way? And *how can the omission of a disconfirmation create duties and rights?*

To understand this issue more clearly, suppose that Alice is a Bayesian rational agent, that is, suppose that H_i is her hypothesis that Bob will not take the car tomorrow. H_i is characterized by a subjective probability $p(H_i)$, representing her degree in belief in H_i. Assuming that beliefs are represented by a well-defined additive probability function,[20] her degree of disbelief in H_i is given by $1 - p(H_i)$. We have already supposed that such beliefs are warranted by inductive reasoning in which Alice has acknowledged that there is a pattern governing Bob's behaviour such that on almost every Monday Bob does not take his car or, alternatively, that not taking the car on Monday is his best choice given his desire not be stuck in traffic.

Suppose that given Alice's concern about what Bob will do this Monday, she starts looking for *additional evidence* for her belief that he won't indeed take the car. Assuming, as we have done above, that *it is commonly known that she has decided to rely on him*, she happens to notice that Bob is being silent about the truth of her belief about him.

The observation of silence, from a Bayesian perspective, can be treated as a 'datum' S for determining whether Alice's belief about Bob is true or false. Hence, by applying the Bayes' theorem, her belief can be updated accordingly. Moreover, such an update of H_i must be determined relative to its complement $\neg H_i$ according to the usual formula:

$$\frac{p(H|S)}{p(\neg H|S)} = \frac{p(S|H)}{p(S|\neg H)} \bullet \frac{p(H)}{p(\neg H)}$$

From a Bayesian point of view, Alice is interested in the impact of Bob's silence on her belief that he will not take the car tomorrow. Such an impact amounts to calculating the probability that her belief is true, given that she has observed Bob's silence. To derive this value, she needs to compute the *posterior* (i.e. the odds that H_i is true in light of what is known after the observation of S) that equals the *likelihood ratio* (i.e. the second term from the right representing the information value of S with respect to

[20] See the canonical axioms in Savage (1954).

the truth of H_i) multiplied by the *priors* that H_i and $\neg H_i$ are true before the observation of S. In such an inference, in case the probability of observing S when H_i is true differed from when is not true, the likelihood ratio would be different from 1, and the posterior would also differ. In particular, the 'datum' (i.e. Bob's silence) favours the hypothesis H_i when the posterior odds are greater. This happens when the *conditional probability* of Bob's silence given that Alice's belief about him is true is larger than the conditional probability of Bob's silence given that her belief about Bob is false. In such a case, it is said that the observation of S is *diagnostic* or *confirms* H_i and not $\neg H_i$.

However, silence clearly is *ambivalent* evidence in that there are both reasons for believing that it supports Alice's belief about Bob (if Bob does not want to take the car, he will not inform Alice that he is going to take it) as well reasons to believe that it can disconfirm it (Bob could not reach her in time or he has forgotten about her desire to have the car, or he simply does not care enough about Alice's desires, or he wants to harm her on purpose and so on). Whether the evidence is relatively more confirmatory than not is a contingent matter, and depends on the *ratio* between the known conditional probabilities of observing silence on condition that her belief is true or false. If she is a Bayesian rational agent, she compares these information values before updating her belief.

It is however an empirical fact that one of the "best known and most widely accepted notion[s] of inferential error" (Evans 1989, p. 41) is that human reasoning gives undue weight to evidence that supports one's beliefs while discounting evidence that would tell against it, and this tendency is called *confirmatory bias*.[21] A confirmatory bias can be discovered in many different situations in which one assesses the truthfulness of one's beliefs. However the empirical evidence is particularly vivid when one is both *concerned* with what one believes (the so-called *motivated* confirmation bias) and the evidence one is evaluating is *ambiguous* (i.e. it is partly supportive and partly not without exactly knowing how much it is so). In this kind of situation, there is a very strong tendency to interpret information in ways that are partial to one's beliefs, and in particular, in ways in which the *positive* side of the evidence is overemphasized.

In the case of silence, its ambivalence would not be too much of a problem if silence were not often also *ambiguous* evidence.[22] But, if it is so, then it seems plausible to assume that *there is a psychological tendency to read other's silence as a positive evidence for one's belief*.

If we accept the confirmatory bias it may be suggested that, in contexts where we already entertain the relevant expectation and we observe silence by those that can disconfirm it, we update the belief by adopting the best expectation that could be

[21] See Nickerson (1998) for a review of the relevant psychological literature; see Rabin and Schrag (1999) for a mathematical model, though focussed on a different aspect of the confirmatory bias; see Jones and Sugden (2001) for experimental evidence in a decision making context.

[22] That is, one is *uncertain* on how to assess such ambivalence: whether the positive support to one's hypothesis is more likely than the negative one; see Ellsberg (1961). In the case at hand, ambiguity about the evidential value of silence can be understood as a form of uncertainty about the conditional probabilities of $p(S|H_i)$ and $p(S|\neg H_i)$. That is, the agent does not know what the likelihood ratio is because it is as if he considered as possible, in the present circumstances, more than one distribution of conditional probabilities of observing silence, given that the hypothesis is true or false.

associated with the observed evidence: the one that would confirm the belief already accepted. In other words, for the agent holding the relevant belief, *silence regarding one's expectation about other's behaviour (i.e. the forbearance to disconfirm such belief) means that the others will act as expected.*

To interpret silence in this way, one must expect that the other is not hostile towards oneself, otherwise the negative side of evidence would be maximally relevant. However, as we have argued in the previous section, sharing a value entails such shared expectation. On the background of this shared expectation of reciprocal non-hostility, it is reasonable to consider that the "natural" meaning of silence is confirmatory.[23]

It is now understandable why the closer Alice is to fulfilling her desire that Bob does not take the car, the surer she is that he will not take it. Supposing that she has checked up on him several times before Monday morning, all these instances of Bob's silence have confirmed her belief possibly up to certainty.

So far so good for the expectation that Bob will not take the car becoming firmer (i.e. confirmed). But what about the fact that she also feels *entitled* that he does not take it?

First of all, given that the confirmatory meaning of silence is *salient* for both of them (we can assume that Bob is a confirmatory agent just like Alice), if the agents have mutual expectations about each other's confirmatory tendency, then they will both have reason to believe that Bob's silence is a confirmation that he will act as expected (i.e. in the same way that he is used to doing on a Monday). More precisely, since Alice believes that he knows that he has not disconfirmed a belief she had about him, she has reason to believe that Bob cannot but assent to her interpretation (at least from the perspective of their shared bounded rationality). But the same is true for Bob. If he has reasons to assent to Alice's belief, he has reasons to believe that it is reasonable to believe something in those circumstances, and so he has reason to believe that he has actually confirmed Alice's belief about him. If the shared salience of precedence suffices to justify the commonality of our beliefs in future conformity to a convention (Lewis 1969, pp. 35–36), *the salience of silence might justify a common belief in the occurrence of confirmation*.[24]

One relevant consequence of such common knowledge is that, although at the beginning Bob was just *passively* allowing Alice to believe something about him on the grounds of a regularity in his behaviour, under these conditions of common knowledge of his confirmation, Bob's allowing becomes *active* (i.e. Bob is tacitly confirming that he will behave as usual). But, on the basis of the argument developed in Sect. 3, this is sufficient for the principle of Reliability to apply. Since Bob has confirmed Alice's expectation and reliance, Bob acquires *a duty of reliability* and Alice a corresponding *right to rely*. From this it follows that, in case Bob did not behave as expected, Alice would not be simply disposed to feel anger towards him, but she would be *entitled* to feel it, that is, she will harbour *resentment* proper.

[23] Natural meaning is here used in the sense of Grice (1957).

[24] For the aim of this paper, employing the same strategy that Lewis has used to justify the belief that all the agents will conform to the precedent is sufficient. For a recent analysis of the role of salience in conventions see Sugden (2011)

5 The normative consequences of conventions

We now have all the conceptual resources to offer an account of the normative consequences of conventions.

Consider a population in which a conventional regularity to drive on the right prevails. As argued in Sect. 2, the agents in the population regularly rely on the others to drive on the right: everyone believes that others will drive on the right and acts accordingly, that is, he himself drives on the right. Given that a convention entails a pre-existing agreement in desires for the same end (i.e. the common interest not to collide), the expectation of reciprocal reliance is a reason for everyone to rely on each other so that, in this way, their desire for the means (i.e. to drive on the right) is also in agreement.

Trust, as we have suggested above, is a fundamental non-hostile attitude on the part of the trustor because an agent relies on another agent to carry out an action that stems from his motivation, without any coercion. The reason why each relies on the other when they are parties of a convention is that each one expects the other to rely on themselves in the same, non-hostile, way. Hence, in order to trust, everyone has to assume such non-hostile attitudes in the trustees too. Suppose, then, as we have done in Sect. 3.1, that the agents in the population share a value of non-hostility; this value would, of course, promote the disposition to trust each other in this kind of situation.

Finally, assume, as it has been suggested in Sect. 4, that the agents are characterized by a confirmatory bias, and that they are aware of each other's confirmatory tendency.

Under these two assumptions, each time two or more agents interact with each other in a situation that is governed by a convention, *if they keep silent about the expectation of reciprocal reliance that they mutually know they have,* each of them *tacitly confirms* such expectations about each other, even if they are not grounded in direct experience (e.g. the agents might have never met before).

By tacitly confirming their reciprocal expectations, each actively allows reliance on the truth of these same expectations. As a consequence, applying the principle of Reliability, each also acquires both a *right that the others rely on themselves*, and an *obligation to rely on the others*. Each agent has now a right that the others drive on the right (i.e. has a right to be trusted) and an obligation to drive on the right himself (i.e. an obligation to trust the others).

The normative consequences of conventional regularities are, then, to be explained in relation to the tacit confirmation of these shared expectations of reciprocal reliance for agents sharing the value of reciprocal non-hostility. From this point of view, it can be predicted that failing to conform to a prevailing convention will tend to "evoke unfavourable responses from others" (Lewis 1969, p. 99). For instance, one of these "unfavourable responses" is an emotion like resentment, that is anger entitled by the violation of the shared value of non-hostility, which can motivate behaviours aimed at enforcing conformity to the convention. Correspondingly, those who violate a convention will experience guilty feelings springing from their having let the others down.[25]

[25] On feeling 'guilty' for betraying others' trust see, for instance, Battigalli and Dufwenberg (2007).

By appealing to a moral principle like that of Reliability, this explanation is close in spirit to Lewis' perspective.

However the one defended here has two main advantages. The first stems from the fact that it is not clear why acting "in a way that answers to others' preferences, when they reasonably expect one to do so" would be by itself normatively required, if one is not directly responsible for the elicitation of those expectations. To answer this problem we have individuated the role of trust and reliance in conventional regularities in general, and argued that, when reliance is intentionally induced or actively allowed, it implies peculiar normative consequences. We have, then, clarified how the process of tacit confirmation of the background mutual expectations typical of conventions can make oneself responsible for those expectations, even if one has not directly created them. The second advantage is that by grounding the moral principle in Lewis' dispositional theory of values, we have also suggested a way in which this approach can be 'naturalized', and how intrinsically normative emotions like 'resentment' or 'guilt' might be integrated in a naturalistic framework without losing their characteristic link with normative judgements.[26]

Finally, from the perspective adopted so far, the *moral principle strategy* is enough to explain the normative consequences of conventions. What is, then, the relation with the *agreement strategy*? Is it needed to account for the normativity of conventions?

In the next sections, we address this issue aiming to show that the system of mutual expectations that characterize conventional regularities may *also* be the source of *tacit agreements*, that is, agreements without any communication between the parties. It is argued that agreements are normative for the same reasons that conventions are, and that the integration of the moral principle and the agreement strategies has the advantage of clarifying how interacting with a background of shared mutual expectations can turn *weak behavioural regularities into strong personal relations between the agents*.

6 Agreements without promises

It is natural and correct to view the social institution of promise as a device for making agreements.[27] It is also natural, but wrong, to consider agreements primarily as 'an exchange of conditional promises'.[28] No doubt, such an exchange creates a binding agreement. However, an exchange of 'unconditional' promises is sufficient for creating an agreement that each of the parties will do something, no matter what. Mutual conditional promises are the natural model for contracts, but they hardly account for the general phenomenon of agreements. Moreover, promises are only one possible way to create agreements. Another opportunity is to avail oneself of the suggestion

[26] Sugden (2000, 2004) has advanced the hypothesis that we are naturally avoidant of being the object of others' resentment, and contends that this emotional primitive may explain the normativity that springs from conventions, *without appealing to any mediating moral principle*. However, 'resentment' being 'anger on the assumption of having suffered some wrong' presupposes a normative judgment; see Rawls (1971), Verbeek (2002) and Miceli and Castelfranchi (2009).

[27] See, for instance, Rawls (1955) and Scanlon (1990).

[28] Robins (1984), see also Lewis (1969, pp. 83–88); see Gilbert (1993) for a critique of this view.

of a third party that, if meeting the interests of all, might be jointly accepted. Finally, that an agreement be mutual is also dispensable. Giving permission, for instance, is a way to enter an agreement, that is, when an agent gives permission to another one to do something that he has the power to prevent, there is an agreement between the two that enables the latter agent to carry out some action.

6.1 Agreement: a definition

When there is an agreement between two agents, say Alice and Bob, the *consent* of at least one of them is necessary. When one consents, one is consenting that somebody carry out an action or that something else happens.[29]

For instance, Bob may consent that Alice uses his car or to do something himself, say pick the children from school. It may also happen that Bob both consents to Alice using his car, and that he gives her the key. In all these situations, an agreement is established, and consenting is related to the fulfilment of another agent's goal that one could interfere with. *Consenting* then entails that *an agent intentionally refrains from doing something that may impede another agent's goal fulfilment* (avoiding negative interference) or *that he actively does something that creates favourable conditions for such goal fulfilment* (enacting positive interference).[30] But what is common to these cases is that an agent has *the power to interfere* with another agent's goal fulfilment.[31] When an agent consents to another agent's goal fulfilment, the former does not interfere negatively or does interfere positively with the latter.[32] An agreement then presupposes pre-existing asymmetrical power relations between the agents. When there is an agreement at least one agent that could (has the power to) interfere, is not interfering.

However something stronger is needed to establish an agreement. Though it is true that a satiated lion is 'consenting' that a gazelle safely wander around him, the gazelle does not have his consent to do so, and there is no agreement between them. The gazelle should be ready to flee as soon as the lion manifests any change of mind; she may exploit such temporary loss of interest, but she cannot *rely on* the lion only because the lion happens not to have the goal to interfere with her. On the other hand, such reliance would be much more justified if the lion were able to *intend* not to interfere with the gazelle, because, for some reason, *he has adopted her goal* to wander around safely. Hence, *one's consent* (not just a behaviour that happens to consent) to the fulfilment of a goal of another agent *amounts to the intention not to interfere with another agent's goal fulfilment since and until the other has such a goal*.[33]

However, this unilateral intention based on goal adoption is still not enough.

[29] For a similar analysis of consent see Simmons (1979, p. 75).

[30] On positive and negative 'interference' relations see Castelfranchi (1998).

[31] For an extended analysis of power and power relations see Castelfranchi (2003).

[32] For simplicity, from here on, we mention only the negative interference situation.

[33] The adopted goal (i.e. the goal that the other fulfils her own goal) is a necessary reason supporting the intention of non-interference. However it is usually not a sufficient one. One may want additional reasons for granting one's consent.

Suppose that Alice and Bob are living together and that Bob owns a car. Though the car is his property, Alice may informally not 'acknowledge' the car as Bob's, because she does not consider the matter of who uses the car as entirely up to him: she does not consider this choice as depending on him alone. Alice knows that Bob has the keys, and that he has some sort of legal power to interfere with her use of the car (she could be charged with theft, for instance). Alice also knows that she has Bob's consent, and that she can use the car whenever she wants to, but still she 'contests' this power over her. In this case, though all the conditions above might be true (i.e. Bob has an objective power over Alice, and he has the intention not to interfere with her), there is still no agreement between them. It might be said, that Alice uses the car despite the fact that Bob can (he has the power to) interfere with her. To have a full agreement between A and B, then, there must be also *A's acknowledgement of B's power of interference*.

Consider this variant now. Though the car is owned by Bob, it is Bob who 'rejects' his own power over Alice, as far the use of the car is concerned. Whereas, Alice may consider that it is up to Bob whether she uses the car or not, Bob himself contests this fact. If Alice asks his permission to take the car, Bob replies that she does not have to ask for it, that it is her choice whether to take it or not. In this situation, again, there is no agreement between them about Alice's using the car, because, though Bob has the power of interference, he does not *value* such power.

What is it, then, to value a power, and what relation does it bear with the acknowledgement of power?

'Valuing one's power' is not simply having the goal (i.e. desiring) to use it, because it might indeed happen that Bob used his power over Alice in a moment of sudden anger, and that he later regrets it because he in fact contests any asymmetrical relation with her. Bob does not want to be motivated in this way towards Alice, that is, he does not 'value' his power over her. Hence, *one 'values' one's power when one has the goal to be motivated in the use of one's power by one's own goals* (i.e. when one desires that it depends on one's choice whether to interfere or not). To have an agreement, *an agent has to value that power that he has over another one*: the agent values the fact the he is able and in the condition to interfere with the other.[34]

'Power acknowledgement' might be seen, on the other hand, as the 'acceptance' of such power, that is, the decision to refrain from opposing the exercise of the power over oneself if the other wants to exercise it. Acknowledging the power of another makes manifest one's fundamental non-hostility towards the other: i.e. one's disposition not to pursue something if the other does not want one to pursue it. *Power acknowledgement*, then, *amounts to the intention not to oppose the decision of the other agent.* Both valuing one's power of interference and the acknowledgement of another agent's power of interference are necessary conditions to enter an agreement.

Consider now this additional possibility. Alice wants to use Bob's car tomorrow, she has his consent, and she acknowledges his power in this matter. Notwithstanding this, for safety's sake, Alice books a taxi for the following morning. Suppose that Alice is actually quite sure that Bob will stick to his intention and will act accordingly. But

[34] On valuing and second-order desires see Frankfurt (1971) and Lewis (1989); for a critique see Watson (1975).

still she is worried that something unexpected might happen. Assuming a worst-case scenario, Alice decides not to rely on Bob. If this is the case, that is, if there is no uptake of Bob's consent, no agreement between them has been established. And *uptake* is precisely such *reliance on another agent's consent* (i.e. the intention not to interfere with one's desire fulfilment).[35]

Finally, even granted all the previous conditions, they are jointly not sufficient for having an agreement. Bob may know that Alice has a very important meeting tomorrow, and that she needs the car. To avoid creating any obstacle, Bob decides to refrain from taking the car but he does this without informing Alice. As a consequence she does not rely on his consent though she *would* have done so *were* she aware of it. While Bob's intention of non-interference is present, her ignorance of such intention may push her to book the taxi. Thus, to have an agreement it is necessary to know that another agent has the power of interfering with oneself, and that the other intends not to exercise such power. But, as it is standard in many social interactions, even such first-order knowledge isn't enough to have an agreement. Alice may know this fact, while Bob does not know that she knows it. On this basis, Bob may think she will in the end call the taxi, and so he decides to use the car. And so on for all the levels. In any agreement, then, an *epistemic condition is necessary*: there should be *common knowledge* of the intention to not interfere.

The same reasoning also supports other epistemic conditions. An agreement cannot be in place unless the agent, who is consenting to the other's goal fulfilment, knows about such a goal in the first place. And again this fact must be out in the open, by being common knowledge that an agent has the power to interfere with a goal of another one. The acknowledgment of such power made by the other agent must also be a matter of common knowledge, given that an agreement is basically a way to obtain something one wants, without coercing the other to do so. And, finally, both the valuing of one's power and the uptake of the consent need to be, analogously, a matter of common knowledge between the agents.

We have now all the conceptual resources for offering the following analysis of what an agreement is.

Given two agents A and B, and given that B has a power of interference over A, they have an *agreement* between them if the following five conditions hold:

(1) B intends NOT TO INTERFERE WITH A'S GOAL FULFILMENT —*consent* condition
(2) A intends TO RELY ON B'*s* CONSENT—*uptake* condition
(3) B values his own power, that is, B has the goal to BE MOTIVATED BY HIS OWN GOAL TO EXERCISE HIS POWER OF INTERFERENCE—*valuing one's power* condition
(4) A acknowledges the power of B, that is, A has the goal to REFRAIN FROM PURSUING HIS GOAL IF B DESIRES THAT A SO BEHAVES—*no coercion* condition
(5) All conditions above are common knowledge.

[35] For a more extended analysis of the notion of 'uptake' see Thomson (1990).

An agreement of this sort may be called 'unconditional', in the sense that the consent is not conditioned on a symmetrical consent given by the other agent. On the other hand, an exchange of conditional promises originates a 'conditional' agreement in which each consent is reciprocally conditioned on the other. Contracts, for instance, are instances of conditional agreements.

Finally, it is also evident that there can be agreements without promises, at least if we take a 'promise' seriously.[36] Agreements are particular kinds of social relations between the agents, and a promise is one possible way to establish such relations (see also Sect. 6.2). Other possibilities, such as a mere exchange of a request and an acceptance, or a simple unilateral permission without any request, make it clear that no promise, in the end, is necessary.

6.2 The normativity of agreements and the principle of Reliability

All agreements have normative consequences, even those that are unconditional and established without the interposition of a promise. However, according to the present analysis, an agreement is primarily a social relation characterized by specific motivational and epistemic conditions, and no normative relation has been so far mentioned. How, on this account, is it possible to explain the 'obligation' of the consenting agent, and the corresponding 'right' to do or to have what an agent has been consented to? Or, to put it differently, *what is the wrong of infringing an agreement*?

It should be clear by now that we may understand the normative consequences of agreements by appealing to the *principle of Reliability* introduced in Sect. 3. On this basis, it may be suggested that *an agreement has normative consequences because the consenting agent is either intentionally inducing or actively allowing the other to take up* (i.e. reliance that the consenting agent intends not to interfere with the other's desire fulfilment). From this it follows that the agent taking up acquires a right to rely on the other. But what exactly has he a right to rely on?

As already pointed out, if there is an agreement between Alice and Bob that Alice will take Bob's car tomorrow, Bob's taking it would be something *wrong*. At first approximation, by taking up another's consent, one acquires a right to a certain behaviour: i.e. that the other does not interfere with his desire fulfilment. But it seems to be even more than this: if, without taking the car, Bob is uneasy when Alice takes it, it seems again that Bob is doing something wrong. If an agreement were there only to rule behaviours, Bob's refraining from taking the car should be enough for complying with its terms.

To understand why this is not so, and what the peculiar normativity of an agreement is, consider first the difference between the mere fact that *some agents agree in their desires* and the situation in which *they establish an agreement between them*. When two agents agree in desires "the same world would satisfy the desires of both" (Lewis 1989, p. 119), even without establishing any social relation between them. On the other hand,*establishing an agreement aims precisely to create such agreement in desires*, but facing the fact the things could have been different. Actually, by acknowledging

[36] Not every obligation to act in some way in the future entails a promise; see Scanlon (1990).

the power of another agent over oneself, one also manifests one's desire to refrain from doing an action, if it is against the desire of the agent endowed with the power of interference. Correspondingly, whatever reason an agent has for giving his consent, he is also informing that the fulfilment of such desire 'agrees' with his own desire in the present conditions (i.e. the agent for some reason desires not interfere with the other). Thus, *by establishing an agreement, the agents come to know that their desires agree*: they are co-realizable and they are so without any coercion.

What then, on this basis, is the peculiar normativity of agreements as social relations?

Recall that by giving one's consent, one intentionally induces or actively allows another agent's reliance on that consent, that is, not just on the observable behaviour of non-interference but, more specifically, on the *intention* not to interfere. Moreover, given the details of the social interaction between the agents, it is also manifest that this decision of non-interference is based on the fact that the consenting agent's desire agrees with that of the other. Hence, in order not to be hostile with the consenting agent, *one has to rely on the decision of non-interference that is based on such 'agreeing' desire*. As a consequence, and given the principle of Reliability, those who rely on a consent acquire a *right to such intention of non-interference based on an agreement in desires*. The consenting agent that is willing to enter an agreement is not only obliged not to interfere (i.e. not to take the car himself), he is also bound *not to change his mind*, otherwise the basic non-hostile attitude of the relying agent would be frustrated: the consenting agent is not only bound not to revise his intention but also not to *desire* to interfere with the other. In other words, *when giving one's consent, one is obliged to keep one's desire of non-interference in agreement with the relevant desire of the other agent*. It is for this reason that even the expression of Bob's uneasiness with what Alice has done is illegitimate because such a reaction would inform her that Bob has actually changed his mind.

Moreover, the agent who relies on the consent is bound in the same way. Since the consent is given on the assumption that the other agent has the desire in question (i.e. that Alice desires to use the car), Bob relies on this fact and Alice has induced him to rely on herself. Hence, Alice is bound not to change her desire too. Otherwise, Alice would actually be hostile to Bob, and the opportunity costs he has paid to eventually decide not to interfere with her would then plainly become induced losses.

In conclusion, even in an *unconditional agreement* like this one, *there are reciprocal obligations and reciprocal rights*. By establishing an agreement between them, *the agents become reciprocally obliged and entitled to keep their desires in agreement*. More precisely, the obligation is *to keep one's first-order desires in agreement*. Such an obligation can be seen as a reason for all the parties in the agreement to have a second-order desire that their first-order desires keep motivating their behaviour. Those second-order desires would motivate the agents to do whatever they can to avoid revising their first-order desires.

Does this entail that an exchange of promises has indeed occurred?

The answer is negative. Those who exchange a promise create the expectation that they will do an action in the future *unless the other consents otherwise* (Scanlon 1990). On the other hand, when giving one's consent without the interposition of a promise, a timely warning can still be enough to release oneself from an obligation, because

the other agent has not lost all her valuable alternatives. Those agreements that are not based on promises are simply weaker than agreements based on them. They aim to create and protect desires that agree, and they do so for agents that share the value of not having hostile attitudes.

7 Tacit agreements and the pragmatics of social interaction

Even if agreements can be formed without promises, the easiest pathway to the establishment of the required epistemic conditions is by means of other kinds of speech acts. For instance, in order to consent to Alice fulfilling her desire to have the car, Bob needs to know about such desire in the first place. The most obvious way to ensure this result is for Alice to inform him or to formulate a request. This is often accomplished through *explicit* communication, that is, by conventionally signalling one's desire with language or with a gesture. However, other opportunities are available. For instance, Alice can inform Bob of her desire just by taking the keys of his car because she knows that, since Bob is looking at her, he will infer the goal behind her behaviour. Analogously, by acting in order to remove an obstacle for Alice, Bob can communicate his consent without language, gestures or other conventional means. This is possible because *practical actions (or forbearances) made with a communicative intention* (i.e. practical actions carried out also with the intention that another agent will believe something by 'reading' such behaviour) *might suffice to send a message*.[37] This form of *implicit* communication through practical actions and their effects might support the creation of agreements that can be dubbed, for this reason, *implicit agreements*. When there is an implicit agreement between two agents, the one having the power to interfere with the other can give his consent implicitly, that is through some practical action implying his intention to refrain from interfering, knowing that the other will understand what is happening. Taken from this perspective, the vast majority of agreements that is usually qualified as 'tacit' are instances of agreements established, silently, via implicit communication.

Notwithstanding this, if there are cases in which it is already common knowledge between the agents that one of them wants something, we argue that even implicit communication may be not needed because the consent, the uptake, and all the other conditions can become common knowledge without either explicit or implicit communication. But how is it possible that all the required epistemic conditions be satisfied, without either promises or any form of communication between the parties? Or, in other words, *how is it possible to have agreements without communication?*

7.1 The interactional pragmatics of tacit consent

Though giving one's consent necessarily is the communication of one's intention of non-interference via the usual Gricean mechanism (Grice 1957), one can *have* another agent's consent without it having been given.

[37] This kind of communication may be named 'behavioural implicit communication', see Castelfranchi (2006), Tummolini and Castelfranchi (2007), and Tummolini et al. (2009).

Consider again the example of Alice, Bob and the car.

We have argued so far that when the parties consider silence a confirmation of the beliefs on the truth of which one relies, the confirming agent becomes *obliged* to be reliable, even if no communication has occurred between them. In the example, Bob becomes obliged not to take the car tomorrow because he has tacitly confirmed Alice's belief that he will not take it. As a consequence Bob has actively allowed Alice's reliance. However the mere fact of not taking the car (and as a consequence of not interfering with her) is not in itself sufficient for Alice to have Bob's consent. According to the analysis developed in Sect. 6.1, when one is giving another agent one's consent to something, one has the intention not to interfere with another agent, that is, *the consent condition implies that the content of the intention refers to another agent's desire fulfilment*. However, the intention behind the behaviour that contingently happens not to create obstacles for another agent needs not be social in its content. Actually, in the example, Bob's decision not to take the car this particular Monday is motivated either because of his habit on Mondays, or because it is the best option he has to avoid being stuck in traffic.

On the other hand, as noted in Sect. 3, when the principle of Reliability applies, one incurs a *directed* obligation: e.g. Bob is obliged *towards* Alice not to take the car, and Alice has a right *against* Bob to this behaviour. Therefore, such a directed obligation is not simply to avoid taking the car, but, more precisely, *to refrain from doing what would prevent her desire fulfilment in that context*, which amounts to being obliged to refrain from interfering with Alice's desire that Bob does not take the car.

Granted that Bob is so obliged, on what grounds can Alice infer that he also *intends* not to interfere with her, i.e. that Alice has Bob's consent and not only a *right* to it?

To see why she may be justified in believing that she has his consent, recall that Bob's silence is confirmatory of her belief that he will not take the car this Monday if they share an expectation of reciprocal non-hostility; otherwise the disconfirmatory reading of Bob's silence would be maximally relevant. But if Alice and Bob share such an expectation, then it may be suggested that they will also *presuppose* that the principle of Reliability is not only impinging on Bob but that he will *actually* conform to it.[38]

In order to understand how such a presupposition might come into being, let's first consider Lewis's analysis of the *kinematics of presuppositions* in a conversation (Lewis 1979b, p. 347). According to Lewis, for some things that may be said during a conversation to be acceptable (e.g. to have a truth value), a certain presupposition must be present, and, by saying those things (provided that nobody objects) the required additional proposition becomes presupposed (i.e. believed to be commonly believed) by all the participants in the conversation.[39]

[38] A *pragmatic presupposition* is different from a 'logical' presupposition and is defined as what each participant *believes to be common belief with all the others*; it is a propositional attitude in which one agent "take[s something] for granted, or at least act as if one takes it for granted, as background information – as *common ground* among the participants in the conversation"; see Stalnaker (2002).

[39] "Presuppositions evolve according to a *rule of accommodation* specifying that any presuppositions that are required by what is said straightway come into existence,provided that nobody objects" (Lewis 1979b, p. 347, emphasis added). The classical example is that of saying "The king of France is bald" that requires the presupposition that France has one king, and one only; see for instance Lewis (1979b, p. 339).

Though presuppositions and their dynamics are typically studied in the context of conversation, non-linguistic joint action in general may have the same properties and consequences of linguistic exchange, as clearly stated by Grice himself since the beginning (Grice 1989, p. 28).[40] Thus, adapting the notion of presupposition to the joint action context, if two agents share an expectation of reciprocal non-hostility, and one has actively allowed another agent's reliance, then *what is required for keeping the reciprocal non-hostility belief true comes immediately into existence*. In the example, since Bob has tacitly confirmed Alice's beliefs about his future behaviour, Alice is justified in believing that Bob will in fact conform to the principle of Reliability because this belief is needed for preserving the truthfulness of the belief that Bob is not acting in a hostile way towards her. Hence, *if they share the expectation of reciprocal non-hostility, then, according to Lewis's rule of accommodation, they also come to commonly believe that the principle of Reliability will actually be followed*. To distinguish it from the special case of language use, this kind of presupposition may be called an *interactional presupposition*.

But once interactional presuppositions are allowed, there is also room for extending all kinds of other pragmatic notions to contexts of non-linguistic joint action, such as, for example, 'implicatures'.[41] *In a context of non-linguistic joint action, implicatures may be understood as those components of one's intention in carrying out an action that are not directly controlling the execution of the action*.[42]

Within this view, Bob's action of being silent is interpreted by Alice as a confirmation of the belief that he will not take the car, but *Bob's silence also means* that he intends not to interfere with Alice or, which is the same, that *Alice has Bob's consent*. That is, while the intention of being silent can be seen as the *proximal* intention behind Bob's manifest behaviour, the intention of non-interference with her is the *distal* one. The latter intention is implicated by Bob's silent behaviour, and can be inferred from the tacit confirmation. In fact, it is required that Bob has such an intention in order to preserve the shared presupposition that Bob is not violating the principle of Reliability.

More generally, *if in this context one's silence "naturally" means one's confirmation* (Grice 1957), *the same silence also "implicates" one's consent* (Grice 1989).

Let's take stock. On the background of a shared expectation of reciprocal non-hostility, and thanks to the process of tacit confirmation, Alice can infer that Bob is obliged not to interfere with her. Adapting the rule of accommodation to non-linguistic

[40] "As one of my avowed aims is to see talking as a special case or variety of purposive, indeed rational, behaviour, it may be worth noting that the specific expectations and presumptions connected with at least some of the foregoing maxims have their analogues in the sphere of transactions that are not talk exchanges" (Grice 1989, p. 28). For a more recent development of this generalization to non-linguistic contexts see also Levinson (1995).

[41] *Conversational implicatures* are defined as a component of the meaning in a speaker's utterance which is not part of what is explicitly said. Crucial for implicatures is the fact that the speaker tacitly exploits pragmatic principles, like Grice's Cooperative Principle and maxims, to complete his meaning while relying on the hearer to invoke the same principles for purposes of interpretation; see Horn and Ward (2004). The relation between Grice's Cooperative Principle and the weaker principle of Reliability is left for future work.

[42] For instance, when one turns on the light by flipping the switch, the *proximal* intention controlling the overt movement is 'flip the switch', while 'turn the light on' is the more *distal* one. For the hierarchical distinction between *proximal* and *distal* intentions see Pacherie (2008).

joint action contexts, since they share an expectation of reciprocal non-hostility, they also come to share a *presupposition* that Bob will act according to the principle of Reliability. By sharing this presupposition with Bob, Alice has then reason to believe that Bob actually intends not to interfere with her: by being silent, he *implicates* that she has his consent. Moreover, given that the expectation of non-hostility is shared by the agents, and that both the tacit confirmation and the normative consequences are common knowledge, it is also commonly known that Bob's silence means (implicates) his consent. It is this kind of consent that should properly be dubbed 'tacit' so that *a tacit consent is a consent without any explicit or implicit communication between the parties.* One's consent is tacit when it is *implicated* by something that one is doing (or forbearing to do) against the background of what is already commonly known by the agents. As a consequence, it also becomes commonly known that Alice has Bob's tacit consent without him manifesting it in any way, that is, *without Bob having given it to her*.

7.2 When the agreement is implicated

As we have argued above, another's consent is not enough for establishing an agreement. An agreement between Alice and Bob that he does not take the car this Monday also entails that (1), Bob desires that his actions with Alice are motivated by his desire to use his power over her (the *valuing one's power* condition), that (2) Alice acknowledges this power over her as far as this issue is concerned (i.e. she intends to refrain from pursuing her goal if Bob desires that she so behaves—the *no coercion* condition), and that (3) Alice in fact relies on Bob's consent (the *uptake* condition).

But even these three additional conditions can be established thanks to the same process of interactional pragmatics.

Let's consider first the *valuing one's power* and the *no coercion* conditions.

From what we have argued so far, Bob has not in fact deliberated on whether to give his consent to Alice to something or not: Bob's tacit consent is just implicated by his silence. So, how can such consent be compatible with Bob's valuing his power?[43]

Though the consent is required in order to preserve the truth of the expectation of reciprocal non-hostility, this does not mean that Bob's consent has been coerced or that no other alternative was possible: it is indeed conceivable that if he *had not confirmed* her belief, Alice *would have accepted* his decision to act in ways that interfered with her desire fulfilment. The truth of this counterfactual element, together with the fact that Bob has however confirmed her belief about him, would be sufficient to guarantee that, though he values his own power and she does acknowledge it, she is notwithstanding entitled to fulfil her desire. But how can they know that such a counterfactual is true?

Simply because the expectation of reciprocal non-hostility requires it too.

[43] In other words, this is the same objection raised by Hume against Locke's consent theory of political authority (Hume 1748). Hume has rejected Locke's claim that such authority is the product of a tacit consent whereby "the subjects have tacitly reserved the power of resisting their sovereign" on the account that "an implied consent can only have place, where a man imagines, that the matter depends on his choice", that is, in our words, when a man imagines that by desiring to interfere, he would thereby have successfully exercised his power.

Suppose, on the contrary, that Bob thought differently: he imagines that even in case he hastened to disconfirm Alice's belief, she would have pursued her desire in any case. This belief is incompatible with the truth of his expectation that Alice values non-hostility as much as he does. Given that there was an alternative to what has happened (Bob could have disconfirmed her belief but he didn't) Bob has to infer, if the expectation of reciprocal non-hostility is to remain true, that she would have behaved in a non-hostile way, that is, she would have acknowledged his power. Hence, both the *valuing one's power* and the *no coercion* condition are satisfied because their contents are *implicated* by what is already common knowledge between them. Moreover since both the fact that he is moved by a desire not to interfere with her and that she acknowledges his power are implicated on the background of what they already commonly know, both conditions are common knowledge, or at least potentially so.

Finally, for the social relation between two agents to be an agreement, there should be the *uptake* of the consent and this fact must be common knowledge between the parties. At first glance it may seem that this condition is already established because, in the example, Alice is already relying on Bob not taking the car tomorrow. However, a consent uptake is not just relying on a behaviour that happens not to interfere with one's desire. It is, more specifically, reliance on the other's *intention* not to interfere with such desire fulfilment (see Sect. 6.1). However, in the example, the *consent* condition is satisfied, and therefore Alice also has the opportunity to rely on Bob's intention not to interfere with her, and not simply on his observable behaviour. But how can such uptake on her part be common knowledge between them?

Suppose that she does not rely on his tacit consent. She can do this for, at least, two distinct reasons. She can consider that Bob is not trustworthy enough, in the sense that, though he now desires not to interfere with her, he might change his mind on this issue. Alternatively, she may not want to take his car anymore: it is Alice who might change her mind. However, both these possibilities are incompatible with the mutual expectation of reciprocal non-hostility.

Let's consider the latter case first. If eventually Alice does not desire to take his car anymore (whereas Bob still intends not to interfere with her), Bob will incur losses (i.e. the opportunity costs Bob has already paid) given that he is now relying on the fact that she has this desire. Just as Bob's silence, her silence is also a continuing confirmation of his belief that Alice desires something from him. Thus, Alice too has actively allowed him to rely on something and, as a consequence, he has now acquired a right to the truth of this proposition, for the same reasons discussed above. If it is too late for a warning, she ought to either compensate for his losses or fulfil Bob's expectation, that is, Alice has to keep her desire in agreement with Bob's. Thus, her silence, like his, has in this context a natural or salient meaning: it means a confirmation that Alice still desires what Bob expects her to desire.

On the other hand, given that both agents expect each other to be non-hostile, if, being doubtful about Bob's persistence, Alice fails to rely on his tacit consent, she would incur him losses. And this is something that must be ruled out to preserve the mutual expectation of reciprocal non-hostility. As a consequence, if Alice's silence *naturally* means that she still desires what he expects her to desire, and having common knowledge of the tacit consent, then Alice's silence means also, or better *implicates*,

that she relies on his consent. That is, *even the uptake is implicated in order not to violate the mutual expectation of non-hostility*. Finally, because this fact follows from something they already commonly know, it is again something that they will, or are potentially able to, know in common.

Let's take stock. Though agreements are very often based on communication, there is a kind of agreement that is not based on any form of communication, not even implicit. It is for this kind that we reserve the description of *tacit agreement*. Crucial for the establishment of tacit agreements is the fact that there is a salient interpretation for each other's silence when it is common knowledge that an agent reasonably expects and wants something from another one. Because of the salience of silence as a confirmatory device that we tacitly, and often involuntarily, become *obliged* to be reliable. To account for such normativity the prima facie plausibility of a principle of Reliability has been invoked. Under the assumption that the agents share a value de se of not being moved by hostile attitudes, and that they mutually expect each other to be non-hostile, the agents will also presuppose that the principle of Reliability is actually followed. As a consequence a tacit confirmation also means one's tacit consent, or better, it implicates such consent. Though implicated, such consent is not coerced because it is also implicated that things could have been different, and this counterfactual possibility is a matter of common knowledge. Finally, once an agent has another's consent, the salience of silence, again, guarantees that the last condition for an agreement is satisfied, that is, those who have the tacit consent tacitly confirm that they will not change their mind and, on this basis, they implicate their own uptake.

Tacit agreements are agreements without communication, and are necessarily established by the tacit confirmation of the involved parties. They are *potential* agreements in the sense that there are reasons to believe that all the conditions for an agreement are fulfilled, and this fact is accessible to the parties, at least if they bother to think hard enough (i.e. to infer the appropriate conclusions). Tacit agreements remain potential as long as everything goes smoothly, that is, for example, if the agent who is tacitly consenting, also acts as expected for whatever reason. They become *actualized* agreements when one would like to act against what the tacit agreement mandates, but has to acknowledge that the consent, the uptakes and all the other conditions do actually hold, that is, when he cannot but assent that a real agreement is in place. Finally tacit agreements, as all agreements, create reciprocal *obligations* and *rights* in the parties to keep their desires in agreement, that is, after an agreement is in place, a unilateral change of mind is no longer legitimate.

8 Conventions are sources of tacit unconditional agreements

If, following Hume's suggestion, there is a connection between conventions and agreements, and given that conventions usually persist without the need of communication, the agreements pertinent to conventions must be created without communication, that is, they must be tacit agreements.

Consider a convention to drive on the right sustained by a common interest in avoiding collisions. In Sect. 5, we have argued that, when the agents share a value of

non-hostility and are liable to the confirmatory bias, they will be disposed to read each other's restraint from disconfirming each other's background mutual expectations (i.e. each other's silence) as a *tacit confirmation* of those expectations. But, since they are mutually aware that they are relying on each other, thanks to the principle of Reliability, they will become reciprocally *obliged* to rely on each other and will acquire a *right* to be relied upon.

But granted this, for the same reasons discussed in Sects. 7.1 and 7.2, each agent's silence also *implicates* his or her consent, that is, from the tacit confirmation of their expectations of reciprocal reliance, everyone can infer that everyone else intends not to interfere with their reciprocal desire fulfilment. Given that in a convention, all the agents desire the conformity of all the others, the *tacit consent amounts to the intention not to interfere with the others' desire for one's own conformity*. And since the other's conformity to a convention amounts to the fact that the others do rely on oneself, in the example, one's silence implicates one's tacit consent to all the others that one has decided not to interfere with their desires to rely on oneself. *In a convention, each also tacitly consents to trust the others.*

Moreover, in any convention, conformity is in the individual interest of each agent, that is, everyone trusts the others because it is in the interest of everyone not to collide with the others, and so to rely on the others by driving on the right. Everyone's desire for the means (i.e. driving on the right) autonomously stems from everyone's common motivation not to collide. This very basic capacity (or power) of instrumental rationality is something that everyone *values* and everyone *acknowledges* to the others. If one had known that it was not in the interest of the others to drive on the right, that is, to rely on oneself, one would act accordingly. This is granted both by the fact that the agents are in a coordination problem (see Lewis 1969, pp. 8–24), and in order to preserve the expectation of reciprocal non-hostility.

Finally, each agent's *uptake* of such tacit consent is ensured by tacitly confirming, first, that others' trust on oneself is still something one desires, and, second, by implicating that one does rely on such trust on oneself. That the uptake holds is required again, in order to keep the truth of the expectations of reciprocal non-hostility. The consequence is that each agent does not only trust the others, but also *relies* on the trust on oneself.

Each time the agents, ignorant of each other's identities as they may be, meet and keep silent about each other's mutual expectations of reciprocal reliance, they establish or implicate a tacit agreement to trust each other. Since the tacit agreement is implicated by everyone's silence both as a trustor and as a trustee, the agreement is *reciprocal*: there is a tacit agreement between the interacting agents that they both trust and are trusted by the others. The tacit agreement is *unconditional* because the tacit consents are not conditioned one on the other: they are implicated in relation to the expectations of reciprocal non-hostility, or, in the specific context, the presupposition that the principle of Reliability is in fact followed by everyone. Finally, the normativity of the tacit agreement stemming from a convention is that, by tacitly agreeing to trust each other, *everyone is obliged to keep one's desires for the means in agreement with the others' and has a right that the others do the same*.

9 Why conventions are sources of tacit agreements

A given regularity is a convention for the way it persists, not for its origins: one conforms if the others conform because it is in one's interest to conform. The stability of conventions is guaranteed by this specific motivational structure (i.e. the pre-existing agreement in desiring some end) together with common knowledge of all the conditions specified in Sect. 1. Therefore the individual instrumental rationality of the agents should suffice to stabilize the regularity. Why, then, does a convention have normative consequences? Why, in addition, is it a source of tacit agreements? Isn't it just an additional pressure, made redundant by the reasons the agents already have for acting as they do? Or, to put it differently, what is the role of obligations and rights in conventions?

Though it is true that conventions are stable for the usual self-regarding reasons, *the fundamental condition that ensures stability is that the agents agree in desiring co-realizable ends*. There is no guarantee, however, that they will keep doing so.

After all, a common interest needs not be some ultimate end that we will invariably pursue forever. The ends we agree in desiring are often just means for some further ends we have. All instrumental desires cease to be motivationally effective once the end, in the light of which we pursue some specific means, has been either fulfilled or revised. Suppose, for instance, that Alice and Bob have a common desire to meet each other at least one day during the week, and that they regularly fulfil this desire by sticking to the convention of going to the movies together every Wednesday. Suppose also that Bob is secretly in love with Alice, and hopes that she will fall in love too. But, for Alice, Bob is just a friend that she is keen to meet, and nothing more. This Wednesday, Bob finally realizes how desperate his situation is, how impossible it is that his love will ever be reciprocated, and he gives up his plan to seduce Alice altogether. As soon as he revises his end to meet with Alice, there would no motive at all to keep pursuing the means of going to the movies with her. But, still, by not showing up, Bob would do something wrong and against Alice, something that, notwithstanding his feelings, he may not wish to be moved to do.

In other words, since all the parties to a convention conform (trust) on the assumption of the others' trust, the agents need to be protected against the mutability of the others' interests, a prospect that might compromise each individual project. Since the kind of common interest presupposed by a convention may be as volatile as any other end we pursue, everyone would be at risk if everyone were free to change their mind without taking into account the others in any way. Within this view, *the normative consequences of conventions and agreements act as further assurances, in case one were to change one's desires, by entitling possible influencing actions* (e.g. punishment by reproach), which can motivate the others beside their current desires.

Conventions tend to reproduce agreement in desiring arbitrary means from a pre-existing agreement in desires for the ends. However, by also being sources of tacit agreements between the agents, the arbitrary means are turned into ends to be pursued for themselves, unless one is able to warn the other in time or is prepared to compensate for the others' possible losses.

10 Conclusion

In his paper on causation, Lewis suggests that Hume has defined a causal succession "twice over" (1973, p. 556).[44] Here we suggest that something similar has occurred when Hume defined a convention as:

> a general sense of common interest, which sense all the members of society express to one another, and which induces them to regulate their conduct by certain rules. […] When this common sense of interest is mutually expressed, and is known to both, it produces a suitable resolution and behaviour. And this may properly enough be *called a convention or agreement betwixt us, though without the interposition of a promise*; since the actions of each of us have a reference to those of the other, and are performed upon the supposition, that something is to be performed on the other part (Hume 1740, Book 3, Part 2, Sect. 2, emphasis added).

That convention can be seen as tacit agreements has been often suggested, and is considered as tantamount to the analysis offered by Lewis. But what Lewis has shown is that, in certain conditions, an agreement in desires for the means might stem from our independent agreement in desires for the ends. However, we have argued that an agreement in desires is not the same as an agreement *between* the agents because the latter, but not the former, is a *social* relation between the agents, an interlocked web of beliefs and intentions. When there is an agreement between the agents there are also some additional normative consequences. Whereas a mere agreement in desires may not have such consequences, a full-fledged agreement always does.

In this paper we have proposed that the normativity of conventions is grounded on the same principle characterizing the normativity of agreements: the principle of Reliability. Moreover when the agents infer a full-fledged agreement between them, they become bound to keep their desires for the means in agreement, and by becoming so bound they are assured that even a stranger will not change his mind without some concern for his fellows.

The agreements that stem from conventions are tacit in the sense that they are implicated by what the agents do (or forbear from doing) and without that any communication between them is necessary. In order for this to be possible, we have offered two substantial hypotheses: (1) there is a salient interpretation, in some contexts, of everyone's silence as confirmatory of the others' expectations (an *epistemic* assumption), and (2) the agents share a value of not being motivated by hostile attitudes (a *motivational* assumption). We have also argued that once these assumptions are accepted, they will presuppose that, during their social encounters, the principle of Reliability is in fact followed. If the former hypothesis is compatible with many available empirical data about human decision-making (Sect. 4), the plausibility of the latter is matter for future research.

[44] Hume defines a causal succession both as a succession that institutes a regularity and by way of a counterfactual analysis. The two notions are to be kept distinct, see Lewis (1973).

References

Bacharach, M., & Gambetta, D. (2001). Trust in signs. In K. Cook (Ed.), *Trust in Society* (pp. 148–184). New York: Russell Sage Foundation.

Battigalli, P., & Dufwenberg, M. (2007). Guilt in games. *American Economic Review, 97*, 170–176.

Castelfranchi, C. (1998). Modelling social action for AI agents. *Artificial Intelligence, 103*, 157–182.

Castelfranchi, C. (2003). The micro-macro constitution of power. *Protosociology, 18–19*, 208–265.

Castelfranchi, C. (2006). From conversation to interaction via behavioural communication. In S. Bagnara & G. Crampton-Smith (Eds.), *Theories and practice in interaction design* (pp. 157–179). Hillsdale, NJ: Erlbaum.

Castelfranchi, C., & Falcone, R. (2010). *Trust theory: A socio-cognitive and computational model*. New York: Wiley.

Conte, R., & Castelfranchi, C. (1995). *Cognitive and social action*. London: UCL Press.

Ellsberg, D. (1961). Risk, ambiguity, and the savage axioms. *The Quarterly Journal of Economics, 75*(4), 643–669.

Evans, J. St. B. T. (1989). *Bias in human reasoning: Causes and consequences*. Hillsdale, NJ: Erlbaum.

Frankfurt, H. (1971). Freedom of the will and the concept of a person. *Journal of Philosophy, 68*, 5–20.

Gilbert, M. (1989). *On social facts*. London: Routledge and Kegan Paul (Reprinted 1992, Princeton University Press, Princeton).

Gilbert, M. (1993). Is an agreement an exchange of promises?. *The Journal of Philosophy, 54*(12), 627–649.

Gilbert, M. (2004). Scanlon on promissory obligation: The problem of promisees' rights. *The Journal of Philosophy, 101*(2), 83–109.

Gilbert, M. (2008). Social convention revisited. *Topoi, 27*, 5–16.

Grice, P. (1957). Meaning. *Philosophical Review, 66*, 377–388.

Grice, P. (1989). *Studies in the ways of words*. Cambridge, MA: Harvard University Press.

Hare, B. (2007). From nonhuman to human mind: What changed and why. *Current Directions in Psychological Science, 16*, 60–64.

Henning, H., & Krogh, C. (1995). Obligations directed from bearers to counterparts. In *Proceedings of the fifth international conference on artificial intelligence and law* (pp. 210–218). College Park, MD: ACM Press.

Hohfeld, W. N. (1996). *Fundamental legal conceptions as applied in judicial reasoning and other legal essays*. New Haven, CT: Yale University Press.

Holton, R. (1994). Deciding to trust, coming to believe. *Australasian Journal of Philosophy, 72*, 63–76.

Horn, L. R., & Ward, G. (Eds.). (2004). *The handbook of pragmatics*. Oxford: Blackwell Publishing.

Hume, D. (1740). In L. A. Selby-Bigge (Ed.), *A treatise of human nature* (2nd ed.). Oxford: Clarendon Press, 1978.

Hume, D. (1748). Of social contract. In E. F. Miller (Ed.), *Essays, moral, political, literary*. Indianapolis: Liberty Classics, 1985.

Jones, A. (1983). *Communication and meaning: An essay in applied modal logic. Synthese library* (Vol. 168). Dordrecht: Reidel.

Jones, M., & Sugden, R. (2001). Positive confirmation bias in the acquisition of information. *Theory and Decision, 50*, 59–99.

Levinson, S. C. (1995). Interactional biases in human thinking. In E. Goody (Ed.), *Social intelligence and interaction* (pp. 221–260). Cambridge: Cambridge University Press.

Lewis, D. (1969). *Convention: A philosophical study*. Cambridge, MA: Harvard University Press.

Lewis, D. (1973). Causation. *Journal of Philosophy, 70*, 556–567.

Lewis, D. (1975). Languages and language. In K. Gunderstone (Ed.), *Minnesota studies in the philosophy of science* (vol. VII). University of Minnesota Press (reprinted in his *Philosophical papers*, Vol. 1, pp. 163–188).

Lewis, D. (1979a). Attitudes de dicto and de se. *The Philosophical Review, 88*(4), 513–543.

Lewis, D. (1979b). Scorekeeping in a language game. *Journal of Philosophical Logic, 8*, 339–359.

Lewis, D. (1989). Dispositional theories of value. *Proceedings of the Aristotelian Society, Supplementary Volumes, 63*, 113–137.

Marmor, A. (1996). On convention. *Synthese, 107*, 349–371.

Miceli, M., & Castelfranchi, C. (2002). The mind and the future. The (negative) power of expectations. *Theory & Psychology, 12*(3), 335–366.

Miceli, M., & Castelfranchi, C. (2009). The cognitive-motivational compound of emotional experience. *Emotion Review, 1*(3), 223–231.
Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology, 2*(2), 175–220.
Pacherie, E. (2008). The phenomenology of action: A conceptual framework. *Cognition, 107*, 179–217.
Postema, G. T. (1982). Coordination and convention at the foundation of law. *The Journal of Legal Studies, 11*(1), 165–203.
Rabin, M., & Schrag, J. L. (1999). First impressions matter: A model of confirmatory bias. *Quarterly Journal of Economics, 114*, 37–82.
Rawls, J. (1955). Two concepts of rules. *The Philosophical Review, 64*, 3–32.
Rawls, J. (1971). *A theory of justice*. Cambridge, MA: Harvard University Press.
Robins, M. (1984). *Promising, intending and moral autonomy*. Cambridge: Cambridge University Press.
Savage, L. J. (1954). *The foundations of statistics*. New York: Wiley.
Scanlon, T. (1982). Contractualism and utilitarianism. In A. Sen & B. Williams (Eds.), *Utilitarianism and beyond* (pp. 103–128). Cambridge: Cambridge University Press.
Scanlon, T. (1990). Promises and practices. *Philosophy and Public Affairs, 19*, 199–226.
Scanlon, T. (1998). *What we owe to each other*. Cambridge, MA: Harvard University Press.
Simmons, A. J. (1979). *Moral principles and political obligations*. Princeton: Princeton University Press.
Stalnaker, R. (2002). Common ground. *Linguistics and Philosophy, 25*, 701–721.
Sugden, R. (2000). The motivating power of expectations. In J. Nida-Rümelin & W. Spohn (Eds.), *Rationality, rules and structure* (pp. 103–129). Dordrecht: Kluwer.
Sugden, R. (2004). *Economics of rights, co-operation and welfare* (2nd ed.). Basingstoke: Palgrave Macmillan.
Sugden, R. (2011). Salience, inductive reasoning and the emergence of conventions. *Journal of Economic Behavior and Organization, 79*, 35–47.
Thomson, J. J. (1990). *The realm of rights*. Cambridge, MA: Harvard University Press.
Tummolini, L. (Ed.). (2008). Convention: An interdisciplinary study. *Special issue of Topoi: An International Review of Philosophy, 27*(1–2), 1–164.
Tummolini, L., & Castelfranchi, C. (2007). Trace signals: The meanings of stigmergy. In D. Weyns, V. Parunak, & F. Michel (Eds.), *Environments for multi-agent systems III, number 4389 in Lecture notes in artificial intelligence* (pp. 141–156). Berlin: Springer.
Tummolini, L., Mirolli, M., & Castelfranchi, C. (2009). Stigmergic cues and their uses in coordination: An evolutionary approach. In A. Uhrmacher & D. Weyns (Eds.), *Multi-agent systems: Simulation and applications*. Boca Raton: CRC Press.
Verbeek, B. (2002). *Instrumental rationality and moral philosophy: An essay on the virtues of cooperation*. Dordrecht: Kluwer.
Verbeek, B. (2008). Conventions and moral norms: The legacy of Lewis. *Topoi, 27*, 73–86.
Watson, G. (1975). Free agency. *Journal of Philosophy, 72*, 205–220.