How we rely on each other:

The perception of commitment in joint activities and communication

Francesca Bonalumi

Submitted to:

Central European University

Department of Cognitive Science

In partial fulfilment of the requirements for the degree of

Doctor of Philosophy in Cognitive Science

Primary supervisor: Christophe Heintz

Secondary supervisor: Gergely Csibra

Doctoral advisors: John Michael & Thom Scott-Phillips

Budapest, Hungary & Vienna, Austria

2022

Declaration of authorship

I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, or which have been accepted for the award of any other degree or diploma at Central European University or any other educational institution, except where due acknowledgment is made in the form of bibliographical reference.

The present thesis includes work that appears in the following papers/manuscripts:

Chapter 1: Bonalumi F, Isella M, Michael J (2019) Cueing implicit commitment. *Review of Philosophy and Psychology 10*, 4, 669-688. doi.org/10.1007/s13164-018-0425-0

Chapter 2: Bonalumi F, Michael J, & Heintz C (2021) Perceiving commitments: When we both know that you're counting on me. *Mind & Language*, 1–23. doi.org/10.1111/mila.12333

Chapter 3: Bonalumi F, Scott-Phillips T, Tacha J, & Heintz C (2020) Commitment and Communication: Are we committed to what we mean, or what we say? *Language and Cognition 12 (2), pp. 360-384*. doi:10.1017/langcog.2020.2

Chapter 4: Bonalumi F*, Mahr JB*, Marie P, & Pouscoulous N. (2022) Beyond the explicit/implicit dichotomy: the pragmatics of accountability and plausible deniability. Retrieved from psyarxiv.com/z2bqt. doi:10.31234/osf.io/z2bqt

Other parts of the thesis will be submitted for publication with the following co-authors:

Chapter 5: Barbora Siposova, Wayne Christensen, John Michael

Chapter 6: Johannes Mahr, Gergely Csibra

Francesca Bonalumi

Acknowledgements

I want first to express my gratitude to Christophe for the immense support, scientific and beyond, I received throughout these years. They were not the easiest, and you always sided with me. Thank you. I was lucky to have you as my supervisor.

I would like to thank Gergő, John, and Thom. I could not go through the difficulties of this PhD without your encouragement, your example, and I am grateful I had the opportunity to work with you.

The staff at our department is special, Ági, Eszter, Andi, Boris, Edit, thank you for making everything so easy. And beyond everyone, Réka: thank you for being our rock, for showing that you care, for always being there.

During these years, I had the privilege to be part of the SoC Lab, and the ACES: being a member of these teams has been a great experience. I want to thank you all, but in particular Mia, and Steve, for the incredible kindness. I received a lot of help from Iuli, Dóri, Zsuzsi, Reni, Eszter, Petra in Budapest, and from Priyanka, Lewis, Sophie in Warwick. And Nicole, above anyone else. Thank you.

Most of this work is a collective enterprise, and I was lucky enough to share it with people I genuinely like. Barbora, Johannes, Nausicaa, Margherita, Ákos: thank you for turning this whole thing from scary to fun.

Beyond the work, this PhD have been the most intense experience because I could get to know some of the most amazing human beings, and I could share with them incredible, exciting, painful, boring, sad, terrifying, intense, meaningless and meaningful moments.

For the fun, the trips, the kama muta, for Balaton and for Louisville, for the jams and the bad (bad) movies, for ducog and the bcccd fun, for the escapes to Belgrade and the hikes, the growling, the baths, the night bike rides and the bad pizza, the demonstrations, for the trullo, the jogs and the climbing, the Yard and the karaoke, the book club, for the piffs and the vaccines, for sharing cushions, for being there when the world lost it, for inspiring me to be a better person.

I want to thank those who were with me before the PhD started, and that still bear with me despite the distance and the changes, in particular my father, Gloria, my support system in Vienna and in Italy. Grazie.

You don't survive a PhD without family, and I owe mine to the two people that have been family in Budapest: Otávio and Simily. I love you.

My thesis is about how people rely on each other. There is one person I could rely on since I have memory, who now took the weight of our lives on her shoulders to allow me to complete this dissertation, and who is also the one person I am proud of sharing my DNA with: my sister Claudia, and to her I dedicate this work.

Abstract

People rely on others often and for many things. Friends rely on each other showing up on time when they meet; colleagues rely on other colleagues to do their part of the job; and, in general, people rely on others living up to their commitments. This phenomenon grounding our social life is as natural as puzzling: relying on each other enables mutually beneficial opportunities, but reliance also makes one vulnerable to the whims of those on which they rely. Why and when do people decide to rely on others? How do they manage to rely on others living up to their commitments, when others may have incentives to behave otherwise? In this thesis, I argue that people rely on others doing something when they perceive others to be committed to it. The perception of commitment is based on various cues, including verbal promises, of course, but also more subtle evidence that the fact that a partner is relying is recognised by the partner who commits.

I will first present a psychological characterisation of the phenomena of committing and relying, suggesting that minimal cues of a commitment initiate a self-reinforcing feedback loop that strengthen the perception of both one's honouring and a partner's relying on such commitment.

Chapter 1 and 2 empirically investigate what are these minimal cues. In Chapter 1, I present a set of studies which reveal two factors that lead to perceiving commitment: the effort put in a joint activity and a shared history of repeated and successful interaction. In Chapter 2, I show that mutual beliefs about partner's reliance are crucially involved when perceiving commitment. Chapter 3 and 4 address the topics of commitment and reliance in the context of communicative interactions. In Chapter 3, I show that people hold communicators accountable for breaking implicit promises when such promises were relied upon. By contrast, when what was communicated was not relied upon, the audience does not hold communicators accountable even if promises were explicitly uttered. In Chapter 4, I present a study showing how partner's reliance has an influence on whether denials of implied contents are plausible or not. Chapter 5 and 6 shifts the focus to the development of a capacity to recognise commitments, and how children react to commitment violations. In Chapter 5 I investigate whether 3-year-old children recognise appropriate motives to break a joint commitment, and whether they manifest appropriate reactions in such cases (when a partner had a moral rather than a selfish motive to break a previous commitment). In Chapter 6, I investigate whether 6-to-7-years-old children discriminate between different sources when holding communicators accountable for their misleading suggestions.

Finally, I present two case studies where the perception of commitment plays a key (and problematic) role: the case of sexual consent, and the case of digital communication. I explain how my findings contribute to explain these phenomena and inform policy.

Contents

Introduction

There is as much as one single individual can achieve, and it is inevitable that people need to rely on each other to navigate the social world and achieve greater goals, if not any. This phenomenon regarding human behaviour is as spontaneous as puzzling. How do people come to rely on others? How do people trust that others will behave in their interests? More than occasionally people find themselves longing for similar goals, or for goals that are interdependent or complementary. In such circumstances, their interests align either in terms of goals or in terms of means how to obtain those goals. Nonetheless, the external conditions can rapidly fluctuate and lead eventually to the emergence of different, conflicting interests. Thus, there is no guarantee that others' interests to achieve, or to support one to achieve these goals will be permanent. If people did not rely on each other, it would be impossible to accomplish joint goals, coordinate with each other, collaborate, exchange relevant information, and build long-term relationships. How do people thus manage to do so? How do people rely on each other, despite a blatant (and potentially fatal) risk of defection or of being deceived?

Any account that aims to address this question must, in one way or another, consider one obvious aspect: that *people do rely on each other*. Despite all the odds, people do manage to collaborate, coordinate, make plans, and communicate with each other. In all these circumstances, people believe that one will act according to plans, or that what they communicate to each other is reliable. People rely on each other, that is to say, people act upon this belief, and change their course of action on the basis of this belief. The mere fact that people do such things so easily suggests that people are able to recognise situations when to rely on each other will be beneficial, and when it will not.

Thus, the broad question of how people trust each other translates in a more accessible question: when do people recognise situations in which it is beneficial to rely on each other? Such dilemma is faced by recipients that, in a communicative interaction, are called to determine whether they can rely on what was communicated by their interlocutor (Sperber et al., 2010); likewise, in a cooperative context, the dilemma is faced by agents that are called to determine whether they can rely on their partner (Heintz et al., 2016). What are the conditions that lead people to believe that other will do what is expected from them? What are the conditions that lead people to believe that others will do their part, or provide reliable information when they communicate?

People believe that others will act or inform favourably because they believe that others are committed to act or inform favourably. To support this belief, which we will call throughout

perception of commitment[1], people gather evidence that such commitments are in place. For instance, imagine two friends, Phoebe and Monica, who planned an evening at the cinema. Imagine that Phoebe receives that very same evening an invitation from Rachel to go to the pub instead. Phoebe has a conflicting incentive to fail to honour the cinema commitment, but Monica will nonetheless trust that Phoebe is not going to abandon the cinema plan. The fact that Phoebe had already purchased the tickets; or that both had mentioned how they like that director; or that Monica had stated that she would not go alone to the cinema. Monica will take these facts as evidence that Phoebe is committed to go to the cinema with her. People are sensitive to factors cueing commitments and reliance.

Social interactions, including joint actions[2], often involve uncertainty about what others will do. As when Rachel invites Phoebe to the pub the night she planned to go to the cinema with Monica, many other social interactions may present a temptation for one agent not to follow through. The uncertainty about Phoebe's behaviour can be tamed by the mere fact that Phoebe committed to Monica to go to the cinema: social interactions are made more predictable and less uncertain thanks to commitments (Michael & Pacherie, 2015; Schelling, 1980). Commitments are a useful tool to reduce uncertainty about others' future behaviour because they stabilise agents' motivation to do X (e.g., Phoebe will have additional motivation to go to the cinema), and partners' expectations that X will occur (e.g., Monica will expect that Phoebe will go to the cinema), proving grounds for rebuking in case of a failure. We will see in the next sections how this works, both from a game-theoretic perspective (pp. 4-8) and from a psychological perspective (pp. 9-15).

Despite some disagreement about to what extent joint action necessarily entails commitments (Bratman, 1992; Gilbert, 2014; Searle, 2010), in many social interactions people do expect at least some minimal commitments being in place (Gomez-Lavin & Rachar, 2019). If we are playing in an orchestra, we expect the instruments to follow whenever the director or the score commands; if we are jamming, we expect the instruments to follow each other on a consistent key; if we ask for indications, we expect to be directed to the location through the shortest or easiest route known by our interlocutor; if we meet along a boulevard and start

[1] Across chapters, I will refer to the perception of commitment also as *commitment attribution* and, following Michael, Sebanz, and Knoblich (2016a), *sense of commitment*.

[2] The notion of joint action is very laden, but I will endorse throughout the broad definition pushed forward by Sebanz, Bekkering and Knoblich that joint action is "any form of social interaction whereby two or more individuals coordinate their actions in space and time to bring about a change in the environment" (Sebanz et al., 2006, p. 70).

walking together, we expect the walk not to be abruptly interrupted. But our expectations go beyond a wishful prediction of others' future behaviours. If an instrument fails to adjust, we will feel entitled to reproach the players—and the more the other instruments are dependent on them, the higher the entitlement. If we misunderstand our interlocutor and start walking in the opposite direction, we expect to be corrected—and the heavier the luggage, the higher the entitlement. If we walk together, we expect that stopping to walk together will be acknowledged—and again the more one changed their course of action because of this walk, the higher the entitlement will be (Gilbert, 1990). We expect that people we interact with are somewhat committed to us not only when the evidence presented to us is explicit (e.g., "Let's meet in front of the cinema at 9.45 pm"), but even in cases when the evidence is minimal: and specifically, this evidence being us relying on them.

My dissertation investigates how *partner's reliance* is decisive in modulating the perception of commitment: the more a partner is relying on something occurring, the more one will be perceived to be committed to that something. The higher a partner's reliance, the more one is expected to live up to it, and thus the more their partner will feel entitled to rebuke them in case of a commitment failure (i.e., failing to honour the commitment).

Commitment is a notion widely exploited in many disciplines, from philosophy of language to game theory, from moral philosophy to social ontology. This (ab)use of the notion is partially due to its great analytical and explanatory power. Without dwelling on the normative considerations that are related to rational and moral commitments (Shpall, 2014; see also Löhr, 2021), we will discuss how commitment can be described both in game theoretical terms (as a solution to a strategic problem), and as a socio-cognitive process. Describing commitment as a solution to strategic problems gives us some insights about the concurrent sequence of psychological events.

Commitment as the solution to a strategic problem

Several theories and models had aimed to explain the evolution of cooperation, and how people systematically make choices that are beneficial for others (Hamilton, 1964; Axelrod, 1984; Nowak & Sigmund, 2005; Baumard et al., 2013; Barclay, 2016). Many social situations can be modelled as strategic games, or strategic problems, and the analysis of the payoff structure of the situation gives insights about how a rational agent would behave, and sometimes predict how humans do behave (Maynard Smith, 1982). Some of these problems, such as the trust game (Berg et al., 1995) or the ultimatum game (Güth et al., 1982) can be described as temptation problems: both agents will benefit if a specific sequence of actions is pursued by both, but one of them has the temptation not to follow through. Commitment helps solving this problem (Akdeniz & Veelen, 2021).

The 'commitment strategy' entails that one agent signals to another agent that they will pursue a specific course of action, even in the face of temptation (Nesse, 2001; Schelling, 1980). Commitment in this sense is an act that one performs to influence another agent's behaviour, and specifically to lead them to trust and rely on them pursuing that course of action. Additionally, by performing such act, one will also gain additional motivation for pursuing that course of action. How do people persuade someone that they will do something that they wouldn't do otherwise? And why would they be perceived to remain motivated to do something that may no longer be in their interest to do?

To distinguish credible from fake commitments requires to be able to properly interpret the cues that one has interests in abiding to their commitment. The interpretation of such cues presupposes certain cognitive skills that we will spell out in the next section (pp. 9-11). For now, we will outline how the credibility of one's commitment depend on the fact that by committing one modifies their payoff structure, making defection less advantageous.

The most (likely) convincing cue to signal such interests is to discard alternative options, and change one's future incentives, in such a way that abiding will remain in one's best interests, or even the only possible choice[3]. For instance, on the way back to Ithaca, Odysseus urges to hear the Sirens' songs but facing the scepticism of his crew who (rightly so) doubt that he (and they) would make it to Ithaca if the songs were listened (Homer, 800 B.C.E./1919). To

[3] Burning bridges, i.e., cutting off any other alternative but the one that they committed to, is allegedly the strongest signal one can convey (Fessler & Quintelier, 2014; Nesse, 2001); for a lab-study showing the long-term advantage of such strategy, see Barclay (2017).

prevent himself from abandoning the crew, Odysseus instructed them to tie him to the mast of the vessel (and to ignore his future orders): such act is one famous and strong example of committing by limiting one's own range of actions.

Most of the time, however, people are persuaded that commitments will be followed through even when incentives are not tangible, and when alternative options cannot be removed. Consider a friendship, for instance, in which both parties rely on the fact that the other will remain loyal even in face of temptations. Emotions are a strong, credible cue of one's interests in abiding to such commitment—and in fact they might even have evolved to serve the function of commitment devices (Frank, 1988; Hirshleifer, 2001).[4]

Formalising a commitment by making explicit promises, oaths, or vows is another way of cueing such interests, and thus increasing the credibility of your commitment. Via such formalised declarations people not only signal their emotional attachment to the commitment, but they also provide evidence of their willingness to put their own reputation at risk—if the commitment is not honoured. The more public the commitment is (and the more people are aware of the commitment), in fact, the higher the reputational stakes will be for the committed individual.

Cueing a commitment by altering your material incentives, by risking your reputation, or by expressing it via emotional displays are similar in one important aspect: if the commitment is credible, the original set of payoffs for performing each action changes (i.e., the expected utility from honouring or failing the commitment) (see Table I). Because one committed, the costs of untying oneself from the mast of a vessel became too high (if not impossible to pay). Similarly, the social costs (e.g., the damage to the self-image and the reputation as reliable individual), as well as the related emotional costs paid in case of a commitment failure (e.g., the disappointment entailed with such failures), are strong incentives to abide to the action that one committed to (Fessler & Quintelier, 2014).

[4] One reason why emotions are perceived to be credible is because they are allegedly hard to fake: some studies suggest that humans are intuitively able to recognise the appropriate emotions when observing a face (Elfenbein & Ambady, 2002), and to some extent also to effectively discriminate between genuine and fake emotional expressions (Ekman et al., 1990; Song et al., 2016).

Table I. *Payoff matrix of the possible choices for the agent who committed.*[5] *When the agent did not commit to A, the expected utility of not doing A depends on the expected reward obtained out of the tempting option; instead, when the agent committed to do A, the expected utility of doing A (i.e., honouring the commitment) is a function of the expected reward obtained out of the committal interaction (assuming that the partner will rely), whereas the expected utility of not doing A (i.e., failing the commitment) is a function of the expected reward obtained out and the expected costs associated with choosing the tempting option.*

	Doing A: Honouring commitment	**Not doing A: Failing commitment**
Not committing to do A	ø	$R_{Temptation}$
Committing to do A	$U_{honour} = R_{interaction}$	$U_{fail} = R_{Temptation} - \boldsymbol{C_{Temptation}}$

In Odysseus' case, if he had not physically committed to going back to Ithaca by tidying himself to the mast, his incentives would have certainly led him to follow the Sirens' call. However, the cost of doing so would be now so high (possibly only tearing his own limbs apart) that his best choice set is finally honouring the commitment. Similarly, in the cinema example Phoebe may have strong incentives to spend the evening watching shows on Netflix. Nonetheless, her previous commitment to go to the cinema with Monica will balance out these incentives. The possibility that Monica may get offended or disappointed is high enough that it doesn't worth the risk. For Phoebe, thus, honouring the commitment becomes more advantageous than failing it.

People are more likely to commit when the expected reward obtained by means of committing (reaching Ithaca; going together to the cinema) is better than the status quo (i.e., $R_{interaction} > 0$). Those who commit are more likely to abide to the commitment when the expected costs that they would incur by giving in to the temptation are higher than the expected reward that they would obtain by giving in to the temptation (i.e., $C_{temptation} > R_{temptations}$). In other words, the costs of ripping off your own's limb are higher than the reward of joining the Sirens; the costs of getting Monica angry are higher than the reward of enjoying Netflix.

The expected utility of the committal interaction for the one who commits is what makes their commitment credible. In fact, by means of credibly cueing such expected utility, the one who commits persuades their audience to rely on this cue. When commitments are credible,

[5] U_{honour}: expected utility of doing A when committed; $R_{interaction}$: expected reward of committal interaction; $R_{Temptation}$: expected reward of tempting option; U_{fail}: expected utility of not doing A when committed; $C_{Temptation}$: expected costs of the tempting option.

the probability that one honour their commitment should increase the probability that their partner will rely on it.

Similarly, the expected utility of the committal interaction for the one who relies is also what lead people to rely on these interactions to occur. As a matter of fact, if the commitment is credible, also the original set of payoffs for relying or not relying changes, namely the expected utility from relying or not relying on the committal interaction to happen will change (see Table II).

Table II. *Payoff matrix of the possible choices for the agent who relied.*[6] *When the agent did not rely on A, their expected utility is a function of the expected reward obtained out of the outside option (irrespective of whether the partner does or does not A); instead, when the agent did rely on A, the expected utility if the other agent does A (i.e., honours the commitment) is a function of the expected reward obtained out of the committal interaction, whereas their expected utility if the other agent does not do A (i.e., fails the commitment) is a function of the expected costs associated with the fact that the other agent chose the tempting option.*

	Partner doing A: Honouring commitment	**Partner not doing A: Fail commitment**
Not relying on partner doing A	$R_{outside}$	$R_{outside}$
Relying on partner doing A	$U_{honour} = R_{interaction}$	$U_{break} = C_{fallout}$

To resume our Netflix-cinema example: Monica, the friend who was asked to go to the cinema, may have had different opportunities to spend the evening, such as going to the theatre or having dinner with the parents. If the expected utility of her outside option is irrelevant, and if the costs she would pay in case Phoebe failed the commitment is failed are inconsistent, it is still advantageous for her to rely on Phoebe's commitment irrespective of whether this commitment will be honoured. But if her outside option is very valuable and the fallout is consistent, the probability that Phoebe will honour the promise to go to the cinema is decisive in shaping her decision of relying on or not relying.

Thus, the expected utility of reliance can be formalised as:

$$[p(\text{Honour})*U_{honour} - p(\text{Fail})*U_{fail}] - R_{outside}$$

[6] U_{honour}: expected utility of rely on A when A is honoured; $R_{interaction}$: expected reward of committal interaction A; $R_{outside}$: expected reward of outside option; U_{fail}: expected utility of rely on A when A is failed; $C_{fallout}$: expected costs of the fallout when A is failed.

That is to say, the expected utility of reliance depends on the probability that the commitment will be honoured or not (and the expected utility that comes out of it), discounted with the secure utility obtainable with the outside option.

The abovementioned description outlines the conditions when it is advantageous to engage in a successful committal interaction (i.e., a situation in which one agent relies on A and the other agent does A). This description leaves us with an important observation: the expected costs of failing the commitment depend on the expected utility of partner's reliance, and the expected utility of partner's reliance depends on the expected costs of failing the commitment. The (probabilistic) fact that the partner relies on A, therefore, must be factored in the expected costs that one will pay if failing to do A; hence in the probability that one will honour the commitment to do A. At the same time, the probabilistic fact that one will honour the commitment to do A similarly affects partner's expected utility of relying on A; hence influencing the probability that the partner will rely on A.

Perceiving people as cooperative utility maximisers

Any discussion about the credibility of one's commitment and the expected utility of one's reliance that do not involve the manipulation of material incentives must take into account how people attribute mental states and preferences to others. Any cue of commitment or reliance will work only if people entertain a set of beliefs about others, i.e., others are perceived to be motivated to possess and fulfil certain preferences. So, when agents perceive others to be committed to, they entertain a certain set of beliefs:

(a) the belief that the other agent prefers to behave in a way that maximises their interests;

(b) the belief that that the other agent has an interest to abiding to their commitment; and, as a consequence,

(c) the belief that the probability that the other agent will abide to the commitment is higher than the probability that they will not abide.

The ability to explain other people's behaviours in terms of their beliefs, goals, preferences, moral dispositions, characters, have been and are still heavily debated (Bermúdez, 2005; Dennett, 1987; Gallese & Goldman, 1998; Leslie et al., 2004). Among the proposals, it has been suggested that the ability to interpret others' actions as intentional is supported by a probabilistic set of inferences, ruled by a 'naïve utility calculus'. According to this principle, people are perceived to behave in a way that maximise their utility. From observable behaviours, it is possible to infer the non-observable causal structure that presumably produced them. This causal structure behind one's action involves one's mental state and preferences—specifically, their evaluation about the costs and rewards they expect to obtain and incur with the (observable) behaviour (Jara-Ettinger et al., 2016). The expected utility from one action or goal is what is perceived to drive people to act and pursue goals.

The idea that people are expected to behave according to their expected utility was borrowed from utilitarians, who first suggested 'utility' (as 'happiness', or 'pleasure') to be the one criterium that should inspire one's action[7]. While utilitarianism is a normative framework about how people ought to behave, the naïve utility theory is a hypothesis in psychology about

[7] According to utilitarians, one ought to act when the overall consequences that would be generated by that act would lead to the greatest good for the greatest number (Bentham, 1789/2007), although qualitatively different pleasures weight differently in bringing about the 'overall utility' (Mill, 1863/2014; see also Sidgwick, 1874/2011; G. E. Moore, 1903/2004). That different pleasures, i.e., different preferences may be contribute to one's expected utility was hinted also by Smith (1759/2006).

people's "intuitive theory" of how others act. This account does not necessarily entail that one's decision-making process is actually governed by such computations: it only entails that people cognise others *as if* their decision-making processes were governed by cost and benefits analyses. This assumption allows to infer different preferences and different motivations on the basis of observable actions.

While involved in social interaction that requires relying on others, people will compute, in probabilistic terms, the likelihood that the partner will honour the commitment to do X (or rely on them to do A): they will compute this likelihood on the basis of the costs and rewards that the partner is perceived to expect by doing or not doing X. The ability to infer other's preferences on the basis of their behaviour is what enables the belief that the other agent will maximise their utility.

We saw how the belief (a) is formed upon the ability to interpret others' actions as not only intentional, but also as driven by a 'naïve utility calculus' (Jara-Ettinger et al., 2016). However, such belief is not enough for a percepton of commitment: abiding to the commitment must be among others' preferences. Thus, the belief (b) takes form upon the supporting evidence, such as tying oneself on the mast, or purchasing cinema tickets in advance. When the evidence is not tangible, however, this belief takes form upon a presumption of cooperativeness, grounded on a general aversion to disappointing others (Battigalli & Dufwenberg, 2007; Heintz et al., 2015).

A vast amount of literature showed without controversy that people manifest prosocial preferences, even when they are costly, and apparently purposeless (Camerer, 2003; Charness & Rabin, 2010; Fehr & Gächter, 2000; Guala, 2012). Among these social preferences, a rudimental one is the preference for not disappointing others. This preference comes along with the assumption that not behaving up to what is expected from us is (socially and emotionally) costly. Experimental evidence shows that participants are willing to pay a cost to avoid disappointing others. For instance, they would pay so that partner is not aware of their role in bringing about an outcome that is good for the participant but less optimal for the partner (Dana et al., 2006, 2007; see also Ockenfels & Werner, 2012). When predicting or being exposed to a partner's expectation about a certain outcome, participants were inclined to choose up to such expectations, even when this entailed a cost for them (Dufwenberg & Gneezy, 2000; Heintz et al., 2015). Being averse to disapponting others' expectations would depend, however, on the kind of expectations that are put on the plate: unreasonable or

unjustified expectations, as well as expectations that are not relied on, do not have the same binding power.

Because people are perceived to have a preference to fulfill others' expectations, commitments will be perceived as credible. By cueing, even only minimally, a preference for doing X, one is providing evidence that they are willing to pay social, emotional, and reputational cost in case X is not followed through; and since people are perceived to be averse to pay such costs, one can presume and trust that commitments will be honoured.

Thus, the fact that people are perceived to be utility maximisers (a) and averse to disappointing others' expectations (b) causes the belief that one's commitment will be honoured, or at least that it is more likely to be honoured than not (c).

The commitment-reliance loop

Describing commitment and reliance as strategic problems left us with the observation that the expected utilities of honouring and relying on a commitment are interdependent. We described also that this interdependency presupposes the recruitment of two psychological attitudes, such as perceiving people as utility maximisers and averse to disappointing others. We will describe now, critically for our purposes, how the commitment-reliance interdependency is rooted in psychological events.

The psychological events at stake are beliefs about the likelihood that a partner is willing to honour or to rely on a commitment X. These probabilistic beliefs, or subjective probabilities, are motivated on the basis on any evidence for X, however minimal. Partner's subjective probability that one is willing to honour their commitment X has consequences on their own willingness to rely on X (and as such on the probability that they will choose to rely); this has a further impact on one's subjective probability that the partner will rely on X, which in turn has consequences on their own willingness to honour their commitment X (and as such on the probability that they will choose to honour X); and again, this has an impact on the partner's subjective probability that the one is willing to honour their commitment X, and so on (see Figure I).

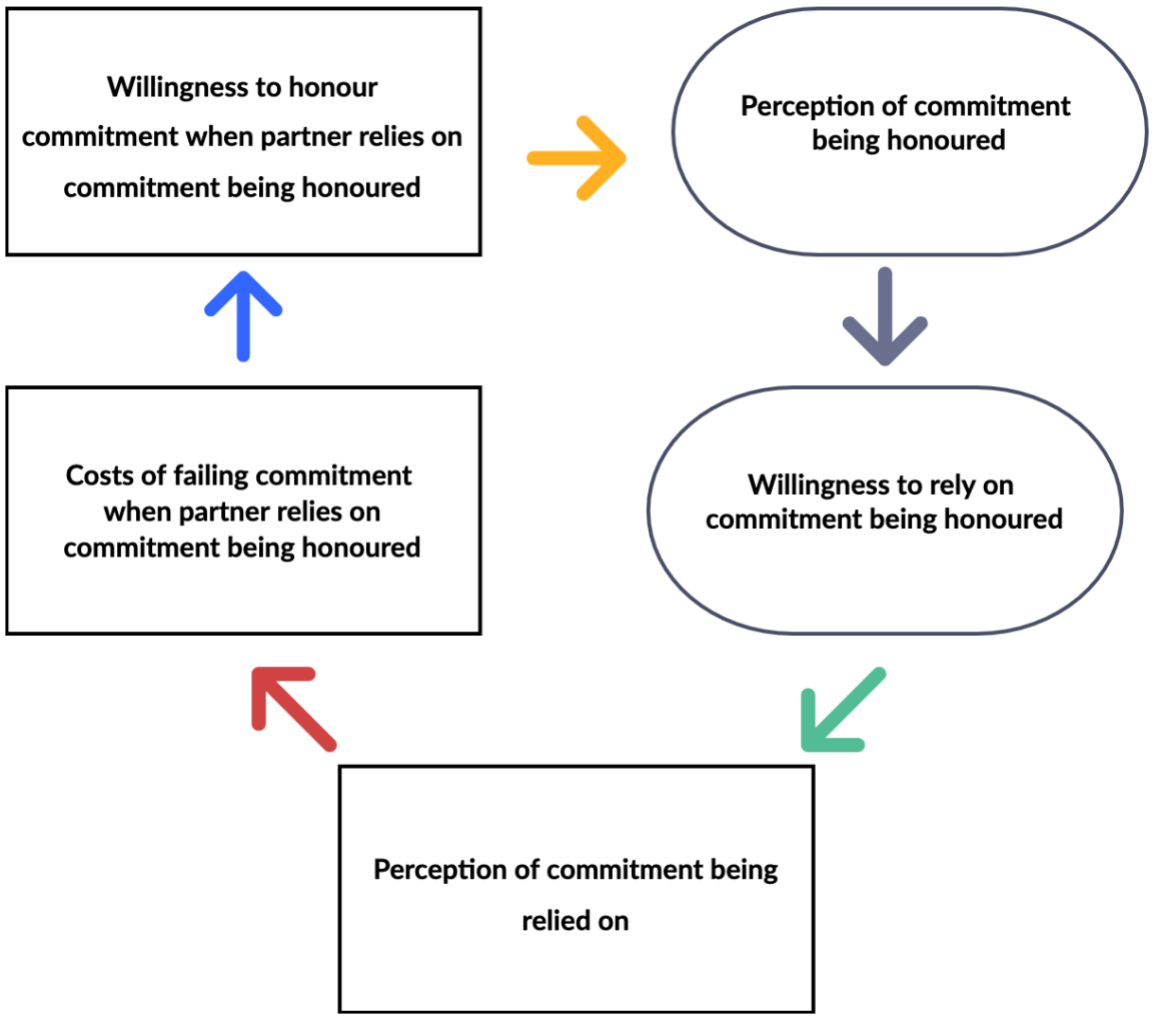

Figure I. A visual representation of the feedback loop involved between partner's reliance, its increasing effect on the costs for the one who commits in case of a commitment failure (and thus its increasing effect on the willingness to honouring the commitment, the subjective probability of the commitment being in fact honour, and the willingness of the partner to rely.

As Figure I illustrates, the commitment-reliance feedback loop unfolds in the following way:

- Perceived partner's reliance, i.e., perception of the commitment being relied on, increases one's costs of yielding to temptation, i.e., costs of the commitment failure (red arrow);
- The costs of the commitment failure increase one's willingness to honour the commitment (increasing the probability that the commitment will be honoured) (blue arrow);

- The perceived willingness to honour the commitment increases partner's subjective probability that the commitment will be honoured (yellow arrow);
- Partner's subjective probability that the commitment will be honoured increases their willingness to rely on the commitment (increasing the probability that the partner will rely on it) (grey arrow);
- The perceived willingness to rely on the commitment increases one's subjective probability that the commitment will be relied on (green arrow).

This dynamic can be illustrated by our previous Netflix-cinema example. If Phoebe thinks that Monica is relying on going to the cinema together, Phoebe will feel more motivated to resist the Netflix temptation and to go to the cinema. This motivation will be even higher if Phoebe thinks that Monica forewent the exciting plan to go out for drink with Rachel *because* she believed that she and Phoebe would go to the cinema together. But if Phoebe thinks that Monica does not really rely on her to go to the cinema, her motivation for staying in and watching Netflix will certainly increase. Similarly, if Monica thinks that Phoebe may decide last-minute to go out with Rachel instead, then Monica will be more motivated to keep her options open for that evening and not rely on the cinema plan happening. Thus, Monica's subjective probability that Phoebe will go to the cinema will reinforce Phoebe's subjective probability that Monica will rely on her going to the cinema. But, on the other hand, Phoebe's subjective probability that Monica will rely on her going to the cinema will reinforce Monica's subjective probability that Phoebe will go to the cinema (see Figure I).

The costs that Phoebe would pay if she gave in to the temptation (i.e., Monica getting disappointed or angry) must be updated in view of Monica's reliance, and these costs are also the evidence provided to Monica about how credible Phoebe's commitment is. Similarly, Monica's and Phoebe's relative beliefs should also be constantly updated. So, the commitment-reliance relation that Phoebe and Monica are involved in will reinforce itself without any need for Phoebe and for Monica to restate their intentions to go to the cinema (although their beliefs may get even more reinforced with such restatements).

The main consequence of the commitment-reliance feedback loop is the fact that, once the loop is initiated, the probabilities of relying on and honouring the commitment (and consequently, the expectation that the other will rely, and the expectation that the other will commit) will tend to reinforce each other—provided that one of the two agents takes action to break the loop and dissolve the commitment. As claimed by Yamagishi and Yamagishi (1998),

as well as by Michael and Pacherie (2015), commitment solves this special kind of temptation problem, and serves its function as uncertainty reduction tool. The reduction of uncertainty brought about by this feedback-loop is what makes commitments so useful in social interaction.

But there is another consequence of this commitment-reliance feedback loop: given that each step of the process will reinforce each other, a committal relation can be cued by very minimal factors and, if no action is taken, still have large consequences in terms of the expected obligations that come out of it. Coordinating to go the cinema may be difficult without the use of verbal agreements, but many other daily life examples suggest that committal interactions are initiated via minimal cues. Even involuntarily: when we open a door without noticing that the person behind us is carrying loads of groceries bags, this action already raises expectations and is perceived to entail obligations, as keeping the door open for them. And even after being explicitly dissolved, such as when lovers agree on 'keep things loose', such interactions continue to raise expectations and are still perceived to entail (to some degree) similar obligations—when honouring and relying on such commitment are still, implicitly, cued.

(Implicit) factors cueing commitment and reliance

The problem of credibility spelled out in the previous section gives credit to the intuition that public and explicit commitments are more 'committal' than non-public and non-explicit ones. The prototypes of such commitments are commissive speech acts—public ones such as oaths, and private ones such as promises. The public and explicit expression of commitment surely counts as a reasonable justification for their audience to rely on the commitment to occur because they increase the costs of defection. The more ritualised, public and explicit such expressions are, the stronger this evidence will justify their partner's reliance on them. Typically, in fact, a job contract, an oath, a wedding, or another public and ritualised statement of agreements and intentions are perceived as very binding, and more binding than similar expressions of agreements and intentions that are not secured in this way.

To confirm this observation, previous research showed that commissive speech acts are predictive of cooperative behaviours in both trust games, and prisoner's dilemmas games (Belot et al., 2010; Charness & Dufwenberg, 2006; Vanberg, 2008). Making expectations explicit, although without them requiring to be confirmed by a commissive speech act, was also found to increase partner's generosity in a modified version of a dictator game (in which recipients can communicate their expectation prior to the dictator's decision) (Heintz et al., 2015), and in a lost wallet game, in which participants were asked to guess the expectations of the first mover (Dufwenberg & Gneezy, 2000). However, such beliefs had a motivating power only when the expectations themselves were deemed reasonable.

Beyond the occurrence of ritualised and public acts, however, people are involved in communicative and social interactions, and they expect and rebuke others when these expectations are frustrated. Although ritualised, public, and explicit acts cue one's commitment and partner's reliance in a straightforward fashion, and thus impact both commitment motivation and perception of commitment, a multitude of non-verbal factors were found to influence these phenomena (as predicted also by Michael et al., 2016a). These factors can be intended or non-intended actions or sets of actions, but also mutually known payoff structures. We claim that these factors influence motivation and perception of commitment by cueing the commitment-reliance feedback loop.

The commitment-reliance loop crosses the boundaries between implicit/explicit, verbal/non-verbal, and intentional/non-intentional. Although typically an explicit speech act cues commitment and justifies more reliance than an implicit one, the loop can be reasonably initiated also by implicated messages, communicative non-verbal actions, and even by non-intended, non-communicative actions, or contextual factors that cue a certain payoff

structure—e.g., how much effort is put in a joint activity may cue one's motivation to engage in the activity, their expectations about the obtainable reward, and their reliance on others doing their part.

Effort turned out to be an important factor shaping commitment motivation. A pioneering paradigm investigating this relation showed that the perception of partner's effort increases persistence in a boring joint task (Székely & Michael, 2018). The study showed that even minimal cue of partner's effort such as the length of the CAPTCHA code solved prior to the joint task had a positive influence on one's persistence in the task itself (a never-ending, increasingly boring, two-person snake game). Critically, this effect was not elicited when participants were told that they were paired with a computer—although the same pattern was found when the partner was reported to be a humanoid robot (Székely et al., 2019), suggesting that perceiving a humanoid robot putting effort into a joint action may cue similar expectations as human agents would. The perception of partner's effort was found to boost not only persistence per se, but also engagement in the task: Chennells and Michael (2018) found that both one's performance, both in terms of time and effort investment, was enhanced when also the partner was perceived to put more effort in it. This study in particular highlights how the commitment-reliance feedback loop works: in fact, your partner's effort put into the task cues both their commitment to completing the task and their reliance on one's contribution; these motivate one's own effort contribution, which in turn justifies partner's reliance on the task being effortfully dealt with, and so on.

Effort is costly, as well as time and coordination. Coordination has been extensively proven to support trust and prosocial behaviour (see for example Atherton et al., 2019; Cross et al., 2016; Kokal et al., 2011; Launay et al., 2013; Wiltermuth & Heath, 2009). One reason why coordination would cue a commitment-reliance relationship is because, in order to enable a coordinated behaviour, agents must implement mutually contingent action plans, which require them to have formed and rely on expectations about the other agent's behaviour: as John Michael notes, "the higher the degree of coordination, the more spatiotemporally exact must those expectations be", and fulfilled (Michael, 2022, pp. 48–49). Participants were found to expect agents to resist to outside options more often when they had coordinated with a partner (Michael et al., 2016b); and to cooperate more, as well as to expect others to cooperate more, when they had act together synchronously (Wiltermuth & Heath, 2009). However, most interpretations of these findings had focused on affiliative feelings that synchrony and coordination would elicit, without taking into account the signalling power that coordinated actions carry. The claim that coordination sustains commitment motivation in view of its

signalling power, and not because of affiliative feelings or overlaps between self/other representations, is corroborated by a study from McEllin, Felber and Michael (2022), which shows that commitment to a partner is boosted by prior coordination only when coordination occurs under conditions of mutual knowledge, and not out of a random sequence of events. In fact, participants invested in the task more effort, and persisted longer in the tasks, when their partner coordinated with them, but only when their action were known by the partner: as such, coordination was a mutually known, intentional outcome. When, however, coordination was achieved without mutual knowledge of each other's actions (making coordination the result of lucky choices), coordination lost its influence on commitment motivation.

Although coordination can be conceptualised as a kind of effort, although possibly a less obvious type, it may have a lesser impact on commitment motivation and perception of commitment, or at it may require some additional element (such as intentionality) to be interpreted as a commitment-reliance cue. Recent findings showed that, while commitment motivation was elicited when effort cues were provided by a humanoid robot (Székely et al., 2019), coordination cues did not work as well, as the effect of coordination on the motivation to resist outside options was present with human partners but not with humanoid robot partners (Vignolo et al., 2019).

However effortless they may be, repeated actions can also serve as basis for forming expectations about others' behaviours in the future. In fact, participants were found to cooperate more when sharing a history of successful interaction in stag hunt games (Rusch & Lütge, 2016) or in coordination games (Guala & Mittone, 2010). Further, participants were found to resist tempting outside options more often when they shared history of successful repeated interaction with their current partner than we they did not (Chennells et al., 2022), and even when alternative and more advantageous options were available and costless (Back, 2010).

A study from Schrift and Parker (2014) showed that even not doing anything[8] can be interpreted as a cue of commitment-reliance: choosing from a set that includes a no-choice (do nothing) option informs individuals that they both prefer the chosen path to other paths and that they consider this path alone to be worth pursuing, an inference that cannot be made in the absence of a no-choice option. Thus, the mutual knowledge of the possibility of not

[8] In a similar (and dangerous) way, silence can be interpreted as an ostensive confirmation of an agreement, particularly if the audience expects to be challenged in case of disagreement. In Part IV I will discuss some problematic consequences of implicit cues of commitment-reliance.

choosing (when the alternative is possible) strengthens individuals' commitment to, and increases their persistence on, their chosen path.

A final cue that is worth mentioning is epistemic authority. Information is constantly transmitted among agents, and even when this information is not about one's interests in doing A or in relying on others doing A, still agents can put their reputation at stake in a similar fashion in order to persuade others about the sincerity of what is communicated. One of the strategies that communicators can exploit is claiming epistemic authority over the information. Appealing to one's expertise or referring to an undoubtable source (Sperber et al., 2010) are certainly strong cues that one is betting on the conveyed information being reliable, but most likely the strongest cue one can provide is having being present when a fact happened—which is, in most legislations, considered as direct evidence in criminal processes[9].

By providing testimony that one was present and could perceive (see) the fact A, a witness is not only providing evidence that will be deemed as credible in support of the claim that A is true and relevant, but they are also betting (socially and materially) that A is, in fact, true and relevant. In juridical matters, the costs that the witness is willing to pay are not only reputational, but also material (perjury is a crime in most legislations). In less institutionalised contexts, such punishments are not envisioned; nonetheless, we take epistemic authority to be both a strong evidence to believe what was communicated, and an evidence to hold other accountable if what was communicated turns out to be not the case (Mahr & Csibra, 2018; see also Mercier, 2017).

Effort, coordination, epistemic authority, time (history), and even refraining from doing something may, at the end of the day, boil down to one single factor: cost. The costs that one agent is perceived to be paying (i.e., investing in an interaction) will be taken as evidence for the commitment-reliance loop.

[9] Note about how instead eye-witnesses are unreliable (Loftus, 1981), but the commitment-reliance loop is so strongly cued that despite all the explicit knowledge about this phenomena, we still consider such testimonies as credible.

Summary of the experimental work/outlook

As outlined in the previous sections, my dissertation investigates how *partner's reliance*, which is the extent to which a social partner's is changing their course of action on the basis that the information is true, or the expected action is performed, is decisive in modulating the perception of commitment. The more a partner is relying on someone doing or communicating something, the more that someone will be expected to be committed to something and the more their partner will be entitled to rebuke them in case of a commitment failure.

In the next chapters I will present experimental evidence in favour of this general hypothesis. I operationalised the perception of commitment in different ways. Across the experiments we operationalised the perception of commitment with accountability judgements (Studies 1 to 4), negative emotional reactions (Studies 1 and 2), the tendency to believe a message (Study 6), protests (Study 5), and partner choices (Studies 1 to 4, Study 6).

In the first part of the dissertation, I will present two sets vignette studies investigating the role of one's reliance on the perception of commitment in cooperative joint activities. In Chapter 1 I will present four experiments which show that minimal cues of one's expectations modulate the perception of commitment, as participants hold agent B accountable the more agent A invested costs in a joint activity and the longer they share a history of repeated interaction. In Chapter 2 I will present further four experiments which show that participants hold one agent B accountable and judge them untrustworthy when B had led (even involuntarily and without any verbal action) another agent A to rely on something that B would then fail to do.

In the second part of the dissertation, I will present two sets of vignette studies investigating how reliance modulates the effectiveness of strategic uses of language. In Chapter 3 I will present three experiments which show that participants would perceive agent B accountable for a promise violation no matter whether this was explicitly uttered or only implied (and critically, this would not occur when an explicit but non-relied on promise is uttered). In Chapter 4 I will present a study showing that strongly implicated promises (e.g., relied on) are also perceived as less plausibly deniable than weakly implicated promises (e.g., less relied on).

In the third part of the dissertation, I will present two studies that investigate how children react to commitment violations in different settings: a cooperative and a communicative setting. In Chapter 5 I will present a study investigating whether 3-year-olds protest less when a puppet defect a joint commitment (i.e., abandon a joint activity) if the puppet faces a

conflicting moral dilemma such as helping another agent in distress). In Chapter 6 I will present an on-line study that investigates the effect of different source claims on how much 6-to-7 years-old children believe a given assertion and how accountable they hold the speaker for the truth of that assertion.

In the last part of the dissertation, I will discuss how the results of the studies presented in the previous chapters have implications in applied debates; specifically, I will discuss how the current debate about consent and digital misinformation need to be empirically informed in order to provide effective safeguard for vulnerable groups.

Part I. Partner's reliance affects the perception of commitment in joint activities

In the introduction (pp. 1-21) we outlined how people rely on each other when they perceive other to be committed to do something that is beneficial. People perceive others to be committed when they are provided *evidence* of one's incentives to honour their commitment. Such evidence can be communicative (verbal or non-verbal) actions, but also non-intended, non-communicative actions, or contextual factors that cue a certain payoff structure. As we suggested, when people perceive others to be committed, they do not merely expect them to honour their commitment. People would have affective reactions, namely they would experience negative emotions associated with a commitment failure. They would also have normative reactions, such as moral disapproval, or a sense of entitlement to rebuke or expect an apology from the agent who failed their commitment.

These normative reactions evoke philosophical analyses about whether certain acts ground normative obligations. Theorists in moral philosophy and social ontology discuss commitment in promises and joint actions from this perspective. The obligations that come out from committing can be expectations-based (as we would tend to agree; see also Scanlon, 1998; MacCormick & Raz, 1972), convention-based (Hume, 1739–1740/2000), or 'joint action'-based (Bratman, 1992; Darwall, 2006; Gilbert, 2014; see also Michael et al., 2016a). There is some disagreement about whether joint action necessarily entails joint commitments, and about whether joint intentions are not reducible to individual ones (see Bratman, 1992), but according to some influential theories the commitments that arise from joint actions are normatively binding because they are not reducible to the each party's individual commitments, hinting to the irreducibility of joint intentions to individual ones (Gilbert, 2009). Joint action has been thus seen as the bedrock of the primitive notion of social commitments.

Conventionalist theories of promises, instead, ground the obligations on the fact that there is such a conventional practice in a defined group as 'you keep your promise': this convention grounds the emergence of a norm for which 'you ought to keep your promise' that enables group coordination and mutual trust (Hume, 1739–1740/2000; Rawls, 1971). Expectation-based theorists hold that a promise is not an act that is conventionally interpreted as a promise, but any ostensive act that reasonably leads a partner to rely on something, and obligations arise upon this reliance (MacCormick & Raz, 1972; Scanlon, 1998). More consistently with Scanlon (1998), we claimed that reliance is not only decisive for the formation of a promise, but more generally it is at the core of the perception of commitment. In the following two chapters I will

present two sets of studies investigating the role of one's reliance on the perception of commitment in joint activities.

We present empirical results from two sets of studies showing what it takes for people to perceive that a commitment is in place. In Chapter 1 I present four experiments which show that minimal cues of one's expectations modulate commitment attribution, as participants hold agent B accountable the more agent A invested costs in a joint activity and the longer they share a history of repeated interaction. In Chapter 2 I present four experiments which show that participants hold one agent B accountable and judge them untrustworthy when B had led (even involuntarily and without any verbal action) another agent A to rely on something that B would then fail to do.

Chapter 1. Non-verbal cues of partner's reliance enhance the perception of commitment

The phenomenon of commitment is a cornerstone of human social life. Commitments make individuals' behavior predictable in the face of fluctuations in their desires and interests, thereby facilitating the planning and coordination of joint actions involving multiple agents (H. H. Clark, 2006; Michael & Pacherie, 2015). Moreover, by stabilizing expectations about individuals' future behavior, commitments can also help to support cooperation. As such, the origin and stability of everyday social exchanges and institutions such as marriage, scientific collaboration, and employment depend upon the credibility of commitments. Speech acts such as promises and vows, as well as complex social institutions such as contracts, allow the creation of explicit commitments – i.e., commitments whose terms are clearly understood and accepted by all parties. But even when commitments are not made explicit, they can nevertheless support the same important social functions. Indeed, philosophers such as Margaret Gilbert and Michael Bratman have recently emphasized the role of implicit commitments in joint actions, based on the idea that joint actions are characterized by the existence of a shared goal – the achievement of which is what all parties implicitly commit to[10] (Bratman 1993; Gilbert 1990). Despite the importance of implicit commitment for distinctively human forms of sociality, it remains unclear how people identify, prioritize and assess their own and others' commitments.

Imagine, for example, that two colleagues, Polly and Pam, are in the habit of meeting and chatting together on the balcony of their office building every afternoon during the coffee break (adapted from Gilbert, 2006). Even if they have never agreed explicitly to engage in this routine, they may over time come to feel much the same as they would if an explicit commitment were in place. As a result, if Pam finds herself confronted with some other important obligation or enticing alternative on one occasion, she may hesitate before breaking the routine she shares with Polly. What factors will influence her judgment as to whether it is acceptable to break with the routine? And what factors will shape Pam's response if Polly does fail to show up? Following Michael et al. (2016a), we hypothesize that people's judgments and attitudes about such situations are governed by a sense of commitment, which is modulated by various cues that another agent expects one to perform a particular action, such as the

[10] With substantial differences: while according to Bratman commitment is not a necessary aspect of shared intentionality, but a characteristic consequence of it, Gilbert holds commitment to be a core aspect of shared intentionality: by sharing a goal, subjects are implicitly agreeing to be part of a plural subject of the shared goal.

history of repeated interaction, and cues that another agent may have invested effort or other costs on the basis of that expectation..

This hypothesis builds upon prior research on the role of expectations and reliance in demanding and motivating prosocial behaviour such as maintaining promises or abiding by tacit rules. MacCormick and Raz (1972) and Scanlon (1998) hold that promises have normative force in situations when the promiser leads the promisee to form certain expectations and to rely about their (the promiser's) future behaviour. In another highly influential contribution made in the context of an analysis of how social practices are established and become self-reinforcing, Lewis introduces the idea of a 'presumptive reason', according to which one ought to fulfil others' preferences when it is the case that one is reasonably expected to do so (1969, pp. 97–98; cf. Bicchieri, 2006). Building upon this idea, Sugden (2000) claims that one is normatively expected to perform a certain course of action X when such a presumptive reason is present, and that one is typically motivated to perform X by means of an aversion to frustrating others' reasonable expectations. Sugden also suggests that this aversion mirrors the emergence of a feeling of resentment towards those who have frustrated one's own expectations.

More recently, some empirical research has begun to test these ideas, specifically to probe the cognitive and motivational mechanisms leading people to feel committed and to act accordingly, and to expect the same of others as well. For example, studies using game-theoretical paradigms have shown that people's expectations have a positive impact on the behaviour of their partners. For instance, Heintz and colleagues (2015) found that participants playing the role of dictator in a dictator game made more prosocial choices when they *explicitly* received information about the recipients' expectations—provided the expectations were reasonable (Cf. also Dana et al., 2006; Ockenfels & Werner, 2012).

However, when there is no explicit information about others' expectations, how can people become aware of them? Addressing this question, Michael et al. (2016a) argue that a partner's investment of effort or other costs in a joint activity may provide an *implicit* cue to that partner's expectations about one's contribution to the joint activity—i.e., if the partner were not expecting one to remain committed and to do one's part, then she would be unlikely to invest effort or other costs. Moreover, a partner's investment of effort also provides a cue that the joint activity is of value to her, implying that she may be particularly disappointed or annoyed if one did not remain committed and do one's part. This line of reasoning is also motivated by previous findings suggesting that the cost invested by one agent in order to allow a partner to obtain rewards has an influence on the choices made by the partner (Charness & Rabin, 2010). More recently, Székely and Michael (2018) also found that the perception of a

partner's investment of effort in a joint activity led participants to remain engaged longer despite increasing boredom.[11] In a 2-player version of the classic snake game which became increasingly boring over the course of each round, participants persisted longer when they were given cues of their partner's highly effortful contribution to the game compared to when they were given cues of a partner's low investment of effort.

While Székely and Michael's (2018) finding is consistent with the hypothesis that the perception of a partner's investment of effort led participants to persist longer out of a sense of commitment, alternative explanations are also possible. For example, the perception of a partner's effort might have led participants to infer that the task was particularly worthwhile. Alternatively, the perception of another agent investing effort may have primed them to exert effort as well, irrespective of any sense of commitment to another agent. To address these open questions, we designed a pair of experiments (Study 1a and Study 1b) to probe participants' normative judgments and affective responses to a scenario in which one agent is relying on a second agent who is presented with a temptation to disengage. However, whereas the abovementioned studies focused on the agent who was presented with the temptation (i.e., they were investigating the effect of a sense of commitment upon this agent's motivation), we opted to focus on the other side of the relation. Consistent with Michael and colleagues' hypothesis, as well as with Gilbert's account (1990, 2014), the perception of a commitment being in place implies that while one agent feels motivated to do what she committed to doing, the partner will feel more entitled to expect it to happen, and to blame more the first agent if she fails to do it. We presented participants with vignettes describing a scenario in which one agent had a high degree of reliance (generated by investing a higher degree of effort into a joint activity, i.e., the High cost condition) or a low degree of reliance (generated by investing a lower degree of effort, i.e., the Low cost condition), and a second agent failed to remain committed. We operationalised commitment in terms of the degree to which participants made negative normative and non-normative judgments about the second agent's violation.

We reasoned that if participants made more negative normative judgments and reported more negative emotional attitudes in response to the High Cost condition, this would be difficult to account for in terms of the aforementioned alternative explanations of Székely and Michael's (2018) finding. Indeed, the priming of the partner's effort and the value of an action to an agent can imply an emotional reaction but does not in itself imply any obligation that she

[11] Indeed, if it is the case that such cues typically track others' expectations, then people may respond to them by increasing their commitment to joint activities even in cases in which they do not in fact reflect a partner's expectations.

has to any other agent to perform the action. This additional normative measure we added would provide further support for the hypothesis that a partner's investment of effort in a joint activity enhances the perception that a commitment to that joint activity is in place. We opted for operationalising commitment using a 6-point Likert scale for the following reason: if commitment is modulated by cues of another agent's expectations, rather than by a norm-violation per se, we should expect that participants' judgments would vary between conditions in a graded manner rather than in a binary manner.

As a further test of the hypothesis that commitment is modulated by various cues that another agent expects one to perform a particular action, such as the history of repeated interaction, we also carried out a second pair of studies (2a and 2b). Studies 2a and 2b were designed to probe participants' normative evaluations and affective attitudes in response to scenarios in which one agent failed to remain engaged to a joint activity toward which her partner had either a high degree of reliance (due to having shared a long history of repeated interaction; High repetition condition) or a low degree of reliance (due to having shared only a brief history of repeated interaction; Low repetition condition). We reasoned that a long history of repeated interaction is likely to establish a high degree of expectation of continued interaction, and thus the scenario described in the High repetition condition would be likely to elicit more negative normative judgments and emotional responses than the scenario described in the Low repetition condition. This line of reasoning is motivated by previous research showing that cooperation in social dilemmas such as the prisoners' dilemma can be boosted if participants experience a history of successful coordination—i.e., in the context of behavioural economics paradigms such as the stag hunt (Rusch & Lütge, 2016) or a pure coordination game (Guala & Mittone, 2010). Unlike these previous studies, however, the current study focused on the perspective of the agent whose expectation was disappointed. Moreover, our paradigm enabled us to investigate people's attitudes and judgments about everyday scenarios with a high degree of ecological validity.

1.1 Study 1a: Costs and Commitment I

Study 1a was designed to test the hypothesis that the perception that an agent's sense of commitment to an interaction is enhanced by her or his partner's investment in an interaction. To this end, we presented participants with vignettes describing everyday situations in which an implicit commitment between two agents was violated. We operationalised the sense of commitment with a normative measure (i.e., a normative question prompting a moral judgment about whether an apology was appropriate), and with two additional non-normative measures

(i.e., an affective question asking whether the situation triggered a feeling of annoyance, and an indirect question about how much time the participant herself would be willing to invest to honour the implicit commitment in the scenario described in the vignette).

Methods

Participants

We used Amazon M-Turk to implement a web-based paradigm with a between-subjects design. Since each participant gave only one judgment for each test question, we expected a high variability in our dependent variables. We therefore opted for a large sample size: 200 participants (2 conditions, 100 per group). We included data from those participants who had already begun the experiment when M-Turk registered that this number had been reached. Our data set therefore comprised 260 adults (124 in High cost condition and 136 in Low cost condition) using Amazon M-Turk (110 female; M_{age} = 33.62 years, *SD* = 10.53). No participant was discarded, since none failed the comprehension question. Here and in all experiments mentioned in this chapter, the methods used were in accordance with the international ethical requirements of psychological research and approved by the EPKEB (United Ethical Review Committee for Research in Psychology) in Hungary. All participants gave their informed consent by ticking a box prior to the experiment.

Materials and procedure

Participants were asked to read a vignette describing a hypothetical situation involving a repeated joint activity that gets interrupted. Subjects were randomly assigned to one of two between-subjects conditions (High cost, Low cost). We manipulated the magnitude of costs that an agent invested to maintain the joint activity with the other agent. In the High cost condition, the scenario reads as follows:

> *You and Pam used to work in the same office on the 5th floor, until you were moved to a 1st floor office one year ago. Every day for the past three years, you and Pam have spent your afternoon coffee break sitting out on the 5th floor balcony and chatting, though you never agreed to start doing this. After you moved to the new office down on the 1st floor, you nevertheless continued to walk up to the same balcony on the 5th floor to spend the coffee break with Pam, even though the balcony is five flights of stairs up from your new office. The sequence is broken when one day you walk all the way up the five flights of stairs and wait for Pam during the coffee break, but she doesn't turn up.*

In the Low cost condition, the vignette differs insofar as the new office is around the corner rather than down on the first floor (See ***https://osf.io/8hrnu/*** for the full vignette). After reading the vignette, participants were asked to respond to the following questions, which were presented in this order:

- Normative question: "*On a scale from 0 to 5, to what extent would you agree that Pam owes you an apology?*" [*0= Disagree strongly*; *5= Agree strongly*].
- Affective question: "*If Pam did not apologize or offer any explanation, how annoyed would you be on a scale from 0 to 5?" [0 = not at all annoyed; 5 = highly annoyed]*
- Comprehension question: "*In the scenario described above, where is it that you and Pam spend the coffee break?" [on the balcony, in the lounge, in the cafeteria]*
- Indirect question: "*Now imagine that you're Pam. The reason why you cannot make it is that, while running an errand in town, you learn that your favorite spa is offering free admission until 4 pm. It is currently 2:30 pm. You would like to write a text message to your colleague back at the office to let her know that you won't be coming today, but you notice that your phone is out of batteries. You plug it in to charge in the car. How long would you be willing to wait in the parking lot for the phone to charge before going in to the spa, in order to be able to send a text message to your colleague?" [not at all, 1 minute, 5 minutes, 10 minutes, 15 minutes, 20 minutes, 25 minutes, 30 minutes]*

The normative question was designed to tap participants' explicit moral evaluations of the scenario. We predicted that they would more strongly agree that an apology was in order in the High cost condition than in the Low cost condition. The affective question was designed to tap participants' more intuitive, emotional reactions to the commitment violation described in the scenario. We predicted that participants would indicate a higher level of annoyance if no apology or explanation were forthcoming in the High cost condition. The comprehension question was designed to filter out participants who had not read the vignette with sufficient care to retain the critical information presented therein. The indirect question was intended to tap participants' appraisal of the commitment indirectly, namely by measuring the opportunity cost they themselves would be willing to pay to uphold the commitment. We predicted that participants would indicate a willingness to wait longer in the High cost condition than in the Low cost condition.

Results

For the normative question, participants gave higher estimates in the High cost condition (*M* = 2.38, *SD* = 1.32, *Mdn* = 3) than in the Low cost condition (*M* = 1.87, *SD* = 1.38, *Mdn* = 2), *t*(258) = 3.007, *p* = .003, Cohen's *d* = 0.37 (small effect size). These results were confirmed by additional nonparametric tests, Mann-Whitney *U* = 6728.500, *p* = .003, *r* = 0.181 (see Figure 1.1).

Similarly, for the affective question, participants gave higher estimates in the High cost condition (*M* = 2.28, *SD* = 1.20, *Mdn* = 2) than in the Low cost condition (*M* = 1.94, *SD* = 1.37, *Mdn* = 2), *t*(258) = 2.121, *p* = .035, Cohen's *d* = 0.26 (small effect size). These results were confirmed by additional nonparametric tests, Mann-Whitney *U* = 7169.000, *p* = .032 (see Figure 1.1).

Responses to the indirect question revealed a numerical difference in the same direction, with participants giving higher estimates in the High cost condition (*M* = 4.01, *SD* = 4.15) than in the Low cost condition (*M* = 3.14, *SD* = 3.47), but this difference did not reach statistical significance, *t*(241) = 1.820, *p* = .070. Levene's test revealed a violation of the equality of variance assumption, *p* = .007.

It is worth noting that responses to both the normative and the affective questions tend to cluster around the middle of the scale rather than towards the two extremes. For the normative question, responses tended to be just below the midpoint both in the Low cost condition (*M* = 1.87, *SD* = 1.38, *Mdn* = 2.00), and in the High cost condition (*M* = 2.38, *SD* = 1.32, *Mdn* = 3.00). For the affective question, responses again tended to be just below the midpoint both in the Low cost condition (*M* = 1.94, *SD* = 1.37, *Mdn* = 2.00) and in the High cost condition (*M* = 2.28, *SD* = 1.20, *Mdn* = 2.00). The findings from Study 1a were consistent with our predictions, providing support for the hypothesis that people's sense of commitment to a joint activity can be enhanced as a function of their partner's investment of effort in the joint activity. In order to ensure that our findings were not due to any incidental features of the scenario, we ran a replication study using a different scenario, and predicted the same pattern of results.

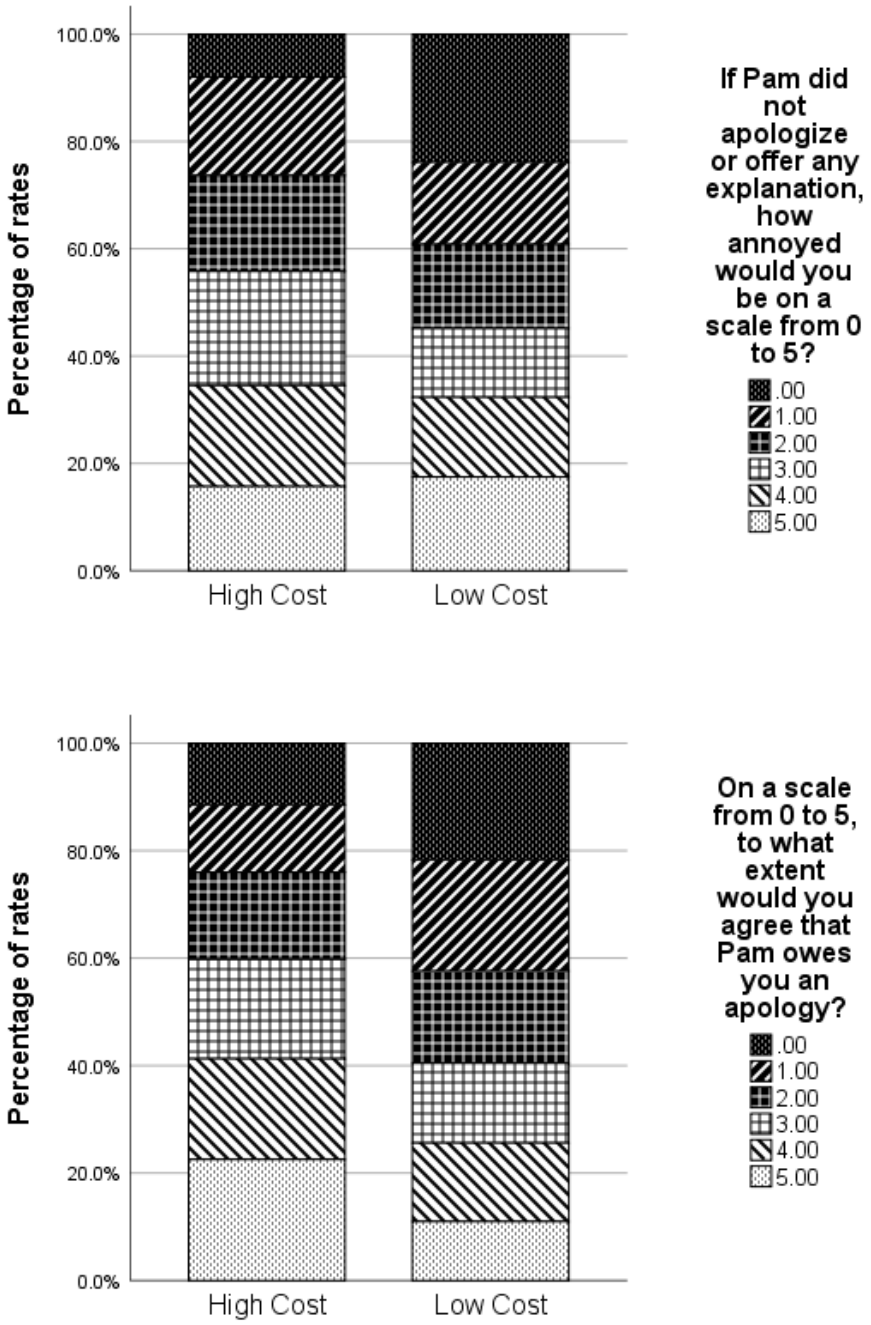

Figure 1.1. Percentage of responses to the normative question (top) and the affective question (bottom). White background bars indicate a mild-to-strong agreement, whereas black background bars indicate a mild-to-strong disagreement with the statement: in other words, the stronger the agreement, the higher the perception of the commitment being violated.

Discussion

The findings from Study 1a were consistent with our predictions, providing support for the hypothesis that people's perception of commitment is enhanced as a function of their partner's investment of effort in the joint activity.

1.2 Study 1b: Costs and Commitment II

Study 1b is a replication of Study 1a and was designed to ensure that our findings were not due to any incidental features of the scenario. For this replication study we used a different scenario and predicted the same pattern of results.

Methods

Participants

As in Study 1a, we used Amazon M-Turk to implement a web-based paradigm with a between-subjects design, aiming for a sample size of 200 participants (2 conditions, 100 per group). We again included data from those participants who had already begun the experiment when M-Turk registered that this number had been reached. Our data set therefore comprised 205 adults. After discarding the data from participants who failed the control question or failed to complete the questionnaire (N = 5), the final sample included 200 participants (105 female; M_{age} = 38.18 years, SD = 11.85), 94 in High cost condition and 106 in Low cost condition.

Materials and procedure

The procedure was identical to the procedure of Study 1a, except that we implemented a different scenario and different control questions. In the High cost condition, the scenarios reads as follow:

> *You and Billy used to live in the same building in the 5th district. Recently, you moved to a different apartment in the 1st district. Every weekday for the past three years, you and Billy have enjoyed jogging together in the park close to your former building, always beginning as soon as the park opens at 7:00 a.m, though you never agreed to start doing this. After moving to the new building, you have continued to join Billy in the same park to jog together, even though the park is on the other side of town from your new apartment. The sequence is broken when one day you wait for Billy but he doesn't turn up.*

In the Low cost condition, the vignette differs insofar as the park is around the corner rather than on the other side of town (See ***https://osf.io/8hrnu/*** for the full vignette).

The questions were presented to the participants in a randomised order, except for the indirect question, which was always presented last.

Results

The results of Study 1a were replicated. For the normative question, participants gave higher estimates in the High cost condition (M = 2.56, SD = 1.46, Mdn = 3) than in the Low cost condition (M = 1.79, SD = 1.39, Mdn = 2), $t(198) = 3.828$, $p < .001$, Cohen's d = 0.54 (medium effect size). These results were confirmed by additional nonparametric tests, Mann-Whitney U = 3484.000, $p < .001$. For the affective question, participants gave higher estimates in the High cost condition (M = 2.50, SD = 1.41, Mdn = 2) than in the Low cost condition (M = 1.68, SD = 1.28, Mdn = 2), $t(198) = 4.317$, $p < .001$, Cohen's d = 0.61 (medium effect size). These results were confirmed by additional nonparametric tests, Mann-Whitney U = 3318.000, $p < .001$ (see Figure 1.2).

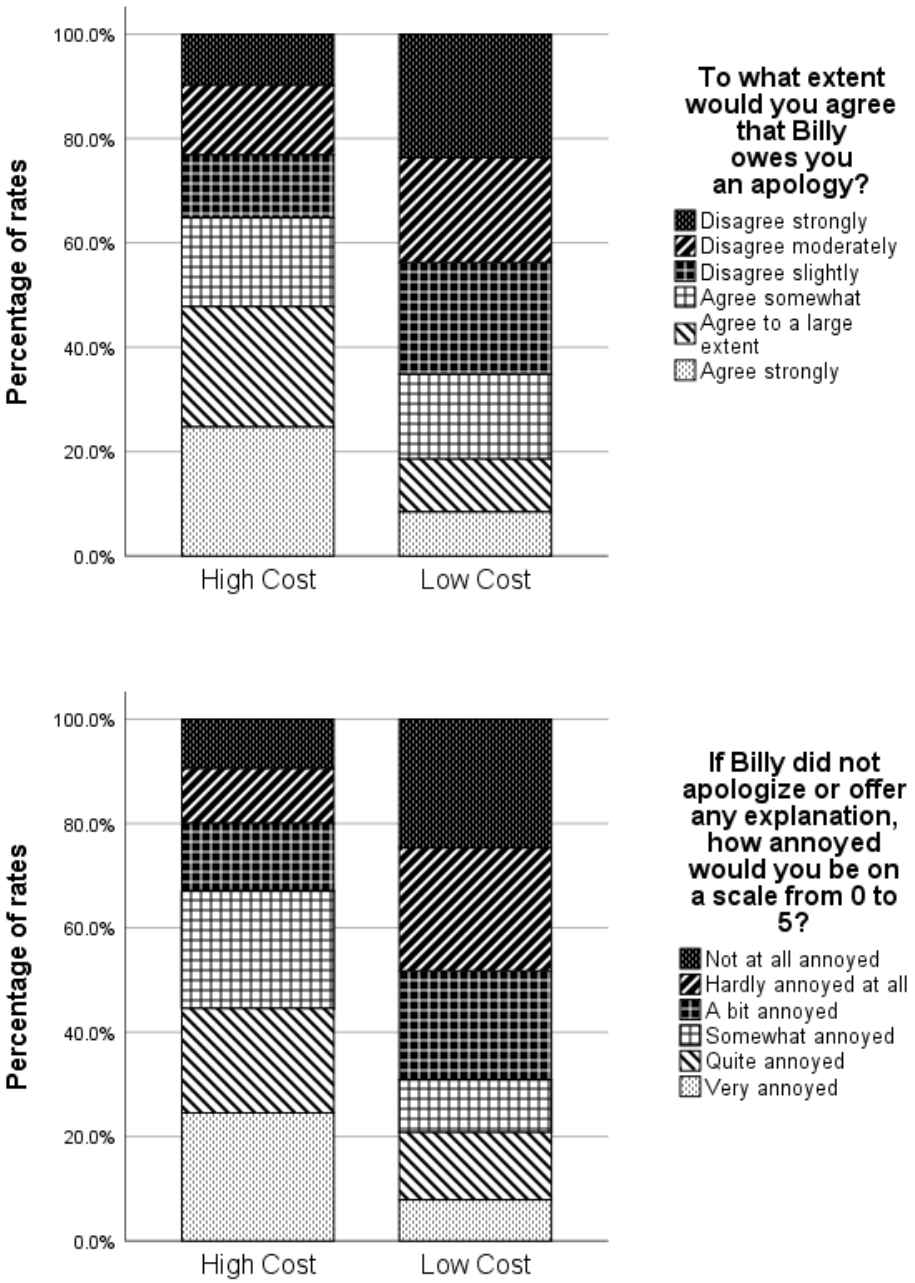

Figure 1.2. Percentage of responses to the normative question (top) and the affective question (bottom). White background bars indicate a mild-to-strong agreement, whereas black background bars indicate a mild-to-strong disagreement with the statement: in other words, the stronger the agreement, the higher the perception that a commitment had been violated.

Responses to the indirect question showed the expected pattern, with participants giving higher estimates in the High cost condition (M = 3.29, SD = 2.73) than in the Low cost condition (M = 3.05, SD = 2.85), but there was again no statistically significant difference between the two conditions, $t(198) = .606$, $p = .545$.

We again found that responses to both the normative question and the affective question tended to cluster around the middle of the scale rather than towards the two extremes (see Fig. 1.3). Indeed, for the normative question, responses tended to be around the midpoint both in the Low cost condition (*M* = 1.81, *SD* = 1.40, *Mdn* = 2.00) and in the High cost condition (*M* = 2.57, *SD*= 1.45, *Mdn* = 3.00). For the affective question, responses tended to be around the midpoint both in the Low cost condition (*M* = 1.70, *SD* = 1.29, *Mdn* = 2.00), and in the High cost condition (*M* = 2.52, *SD* = 1.41, *Mdn* = 2.00).

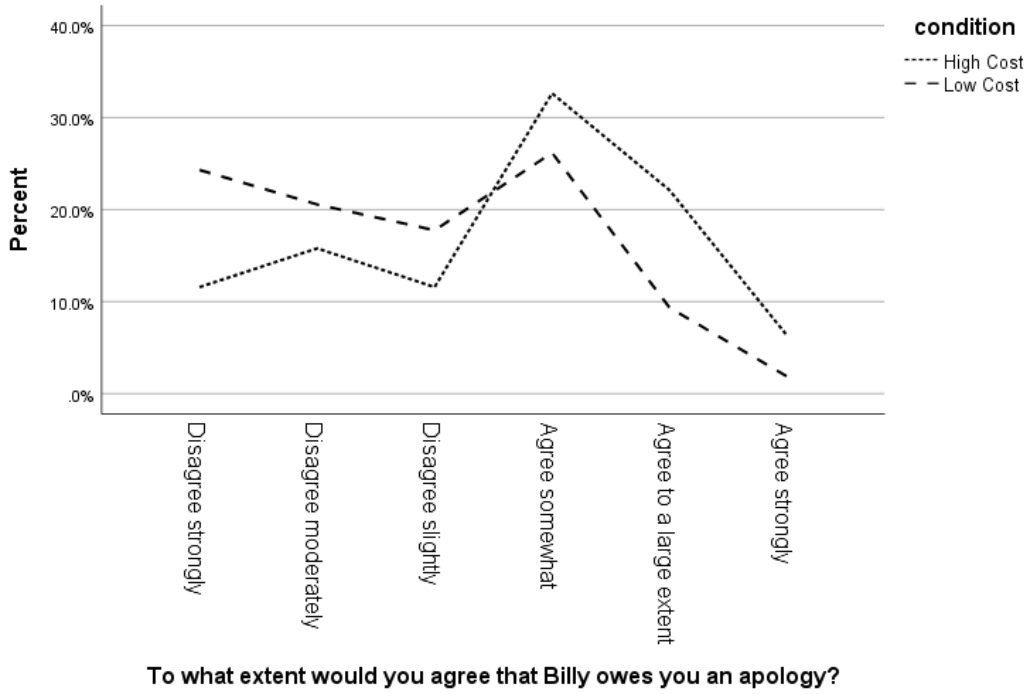

Figure 1.3. Distribution of responses to the normative question. Although in the Low cost condition there is a significantly higher percentage of responses at the lower end of the scale than in the High cost condition, we can see that the largest number of participants in both groups give responses just above the midpoint of the scale.

Discussion

The findings from Study 1b replicate those from Study 1a in a different scenario, which constitutes strong evidence for our hypothesis that one's perception of commitment can be enhanced as a function of her partner's investment of effort in the joint activity.

1.3 Study 1c: Repetition and Commitment I

Study 1c was designed to test the hypothesis that the repetition of a joint activity can enhance people's perception of commitment. To this end, we presented participants with

vignettes describing everyday situations in which an implicit commitment between two agents was violated. We again operationalised commitment attribution with both normative and non-normative measures (i.e., with the normative, the affective and the indirect questions), as we did in Studies 1a and 1b. We marked in bold those parts of the text that implemented the manipulation (i.e., the phrases 'three years' and 'three days'), in order to ensure that participants would not fail to notice these apparently minor details which might be overlooked by a casual reader.

Methods

Participants

As in Studies 1a and 1b, we used Amazon M-Turk to implement a web-based paradigm with a between-subjects design, and again aimed for a sample size of 200 participants (2 conditions, 100 per group). As in the previous studies, we included data from those participants who had already begun the experiment when M-Turk registered that this number had been reached. Our data set therefore comprised 210 adults. After discarding the data from participants who failed one or more control questions (N = 14), the dataset included 196 data from participants, 97 in the High repetition condition and 99 in the Low repetition condition (109 female; M_{age} = 37.74 years, SD = 11.62). The research was carried out in accordance with the international ethical requirements of psychological research and approved by the EPKEB in Hungary. All participants gave their informed consent by ticking a box prior to the experiment.

Materials and procedure

The procedure employed was the same as Studies 1a and 1b. In the High repetition condition, the scenario reads as follows:

> *You and Pam work in the same office building.* ***Every day for the past 3 years****, you and Pam have spent your coffee break sitting out on the balcony and chatting, though you never agreed to start doing this. The sequence is broken when one day you walk up to the balcony and wait for Pam during the coffee break, but she doesn't turn up. This is surprising given that it hasn't happened in the past 3 years.*

In the Low repetition condition, the vignette differs insofar as the coffee break routine was initiated only three days rather than three years earlier (See ***https://osf.io/8hrnu/*** for the full vignette).

Again, participants were asked to respond to normative and non-normative questions. In light of participants' feedback to a pilot version of the study, we opted to introduce a milder normative measure than that used in Studies 1a and 1b. Specifically, we asked participants to evaluate whether the partner who had violated the implicit commitment owed them an explanation (rather than an apology). Also, we opted for an additional question that was tailored to the manipulation of repetition rather than costs—i.e., rather than probing participants' willingness to pay a cost to honour the commitment (as in Studies 1a and 1b), we asked about their willingness to resume the routine. The questions were presented to the participant in the following order:

- Normative question: "*On a scale from 0 to 5, to what extent would you agree that Pam owes you an explanation?". [0= Disagree strongly; 5= Agree strongly]*
- Affective question*: "If Pam did not apologize or offer any explanation, how annoyed would you be on a scale from 0 to 5?" [0 = not at all annoyed; 5 = highly annoyed]*
- Implicit question: *"How interested would you be in spending your coffee break with Pam the next day?". [Not at all interested, Hardly interested at all, A bit interested, Somewhat interested, Quite interested, Highly interested]*
- Comprehension Question: "*In the scenario, where is it that you and Pam spend the coffee break?" [On the balcony, At the cafeteria, In the lounge]*

As in the previous studies, the normative question was designed to tap participants' explicit moral evaluations of the scenario. We predicted that they would more strongly agree that an explanation was in order in the High repetition condition than in the Low repetition condition. The affective question was designed to tap participants' more intuitive, emotional reactions to the commitment violation described in the scenario. We predicted that participants would indicate a higher level of annoyance if no apology or explanation were forthcoming in the High repetition condition. The control question was designed to filter out participants who had not read the vignette with sufficient care to retain the critical information presented therein. The implicit question was intended to tap participants' implicit appraisal of the commitment, namely by measuring their willingness restore the routine if they were in the position of the individual described in the scenario. We reasoned that participants would indicate a lower willingness to restore the routine in the High repetition condition than in the

Low repetition condition, as one's reliance in the former condition would be much higher. If so, then the more severe the violation, the more serious would be the consequences for the violator.

Results

For the normative question, participants gave higher estimates in the High repetition condition (M = 3.13, SD = 1.54) than in the Low repetition condition (M = 2.10, SD = 1.34), $t(189) = 5.014$, $p < .001$, Cohen's d = 0.73 (medium effect size). Since the sample failed Levene's Test for equality of variance (p = .018), we also conducted nonparametric tests, which yielded consistent results, Mann-Whitney U = 2890.000, $p < .001$ (see Figure 1.4).

For the affective question, participants again gave higher estimates in the High repetition condition (M = 3.03, SD = 1.66) than in the Low repetition condition (M = 1.87, SD = 1.17), $t(172) = 5.659$, $p < .001$, Cohen's d = 0.86 (large effect size). Since the sample failed Levene's Test for equality of variance, ($p < .001$), we again conducted nonparametric tests, which again yielded consistent results, Mann-Whitney U = 2841.500, $p < .001$ (see Figure 1.4). These results confirm our prediction, providing support for the hypothesis that a joint activity which has been repeated over a longer period of time elicits a stronger sense of commitment than a joint activity that has been repeated only over a short period of time.

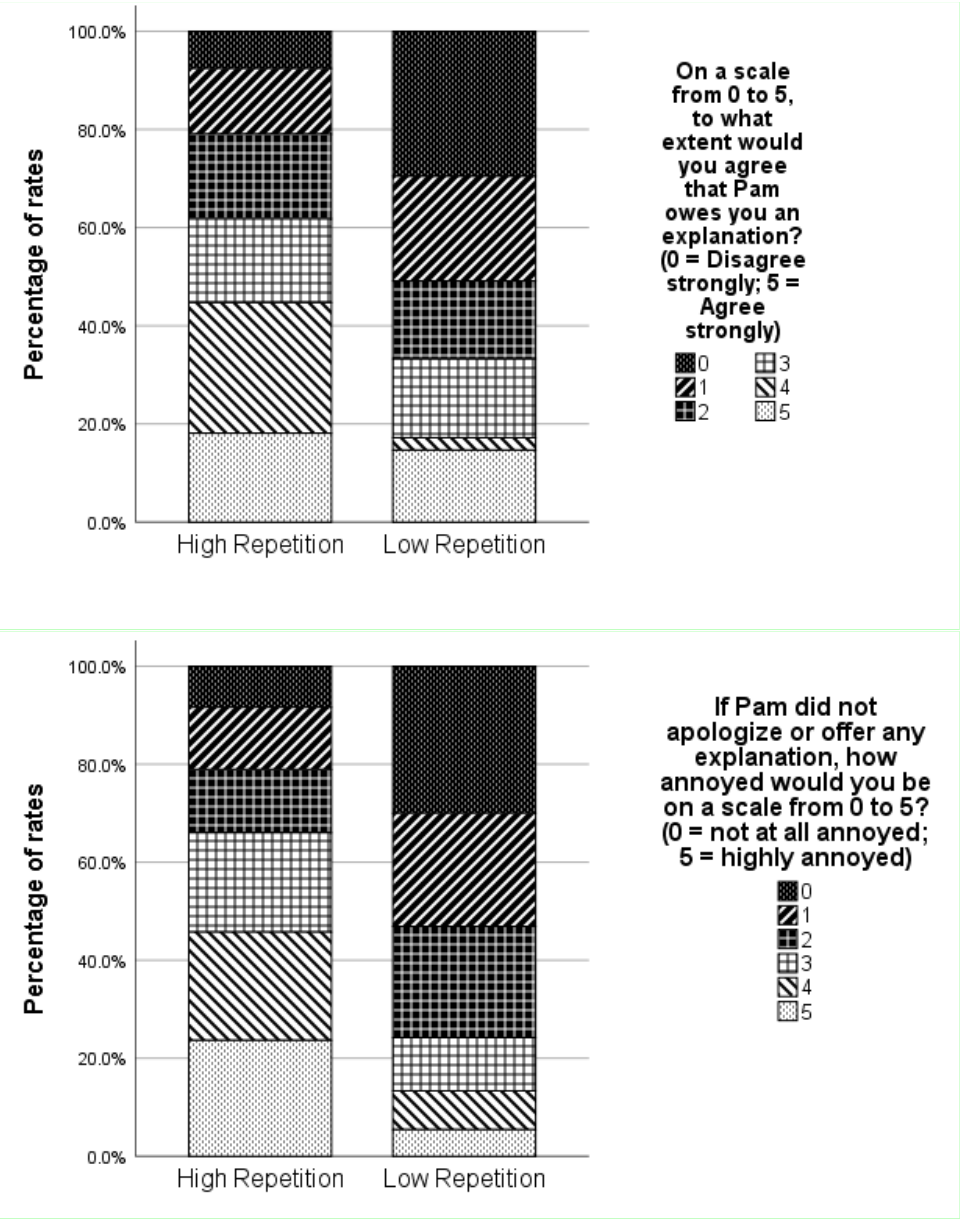

Figure 1.4. Percentages of responses to the normative and to affective question. White background bars indicate a mild-to-strong agreement, whereas black background bars indicate a mild-to-strong disagreement with the statement: in other words, the stronger the agreement, the higher the perception that a commitment has been violated.

The opposite pattern of results was found for responses to the implicit question. Participants indicated a higher degree of willingness to restore the previous routine after a commitment violation in the High repetition condition (M = 5.02, SD = .85) than in the Low repetition condition (M = 4.45, SD = 1.03), $t(189) = 4.186$, $p < .001$, Cohen's d = 0.64 (medium effect size). The sample failed the Levene's Test for equality of variance, p = .005. Nonetheless, this pattern of results is confirmed by a nonparametric test, Mann-Whitney U = 3302.500, $p <$ .001 (see Figure 1.5). Although these findings are not consistent with our prediction, we believe

that they can be explained by hypothesizing that a longer history of interaction gives rise to a more stable sense of commitment, which continues to bind the two partners even after minor violations such as the ones described in both scenarios.

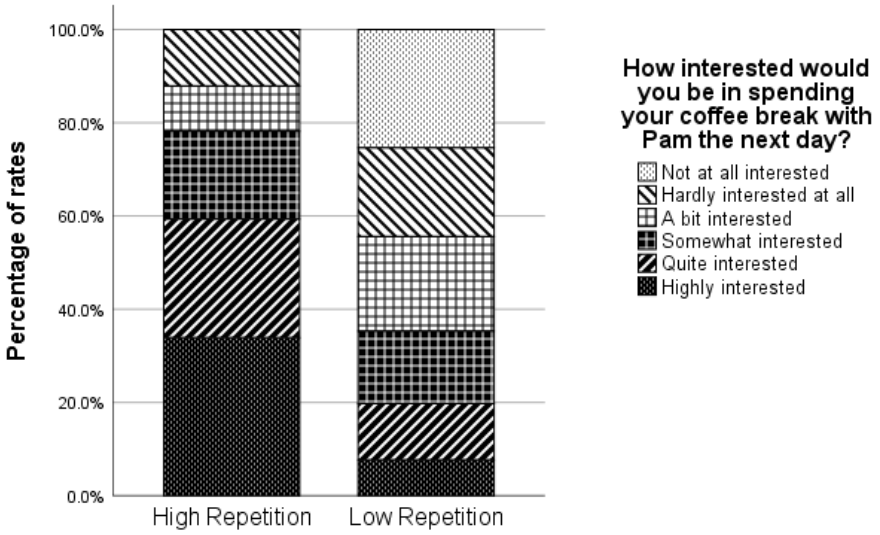

Figure 1.5. Percentage of responses to the implicit question. White background bars indicate a mild-to-strong disinterest, whereas black background bars indicate a mild-to-strong interest in restoring the previous routine: in other words, the stronger the interest, the greater the perception that a commitment is in place.

As in the previous studies, we found that responses did not cluster at the extreme ends of the scale, but tended to be distributed homogeneously across the scale (i.e., distributions were not skewed). For the normative question, responses tended to be right around the midpoint both in the Low cost condition (M = 2.22, SD = 1.40, Mdn = 2.00), and in the High cost condition (M = 3.18, SD = 1.54, Mdn = 3.00). For the affective questions, responses tended to be around the midpoint both in the Low cost condition (M = 2.01, SD = 1.30, Mdn = 2.00), and in the High cost condition (M = 3.07, SD = 1.66, Mdn = 3.00).

Discussion

The findings from Study 1c were consistent with our predictions, providing support for the hypothesis that people's perception of a commitment being in place can be enhanced as a function of their partner's history of engagement in the joint activity.

1.4 Study 1d: Repetition and Commitment II

As previously designed, we ran a replication study with a different scenario, and we predicted the same pattern of results.

Methods

Participants

As in the previous studies, we used Amazon M-Turk to implement a web-based paradigm with a between-subjects design, and again aimed for a sample size of 200 participants (2 conditions, 100 per group). We again included data from those participants who had already begun the experiment when M-Turk registered that this number had been reached. Our data set therefore comprised 203 adults. After discarding the data from participants who failed the comprehension question (N = 12), the sample included 191 participants, 90 in High repetition condition and 101 in Low repetition condition (112 female; M_{age} = 40.49 years, SD = 13.38). The procedure was identical to Study 1c. The research was carried out in accordance with the international ethical requirements of psychological research and approved by the EPKEB in Hungary. All participants gave their informed consent by ticking a box prior to the experiment.

Materials and procedure

The procedure was identical to the procedure of Study 1c, except that we implemented a different scenario and different control questions. In the High repetition condition, the scenario reads as follow:

> *You and Billy live in the same building.* ***Every morning for the past 3 years****, you and Billy have enjoyed jogging together in the park close to your apartment building, each time beginning as soon as the park opens at 7:00 a.m., though you never agreed to start doing this. The sequence is broken for the first time in 3 years when one morning you wait for Billy but he doesn't turn up.*

In the Low cost condition, the vignette differs insofar as the jogging routine was initiated only three days rather than three years earlier (See ***https://osf.io/8hrnu/*** for the full vignette). The questions were presented to the participants in a randomised order.

Results

The results of the previous study were replicated. For the normative question, participants gave higher estimates in the High repetition condition (M = 2.29, SD = 1.51, Mdn = 3) than in the Low repetition condition (M = 1.39, SD = 1.44, Mdn = 1), $t(189) = 4.236$, $p < .001$, Cohen's d = 0.62 (medium effect size). This pattern of result is confirmed by nonparametric tests, both for the normative measure, Mann-Whitney U = 2995.500, $p < .001$ (see Figure 1.6). For the affective question, participants again gave higher estimates in the High repetition condition (M = 1.93, SD = 1.44, Mdn = 2) than in the Low repetition condition (M = 1.29, SD = 1.42, Mdn = 1), $t(189) = 3.110$, $p = .002$, Cohen's d = 0.45 (medium effect size). This pattern of results is confirmed by nonparametric tests, Mann-Whitney U = 3370.500, $p = .002$ (see Figure 1.6).

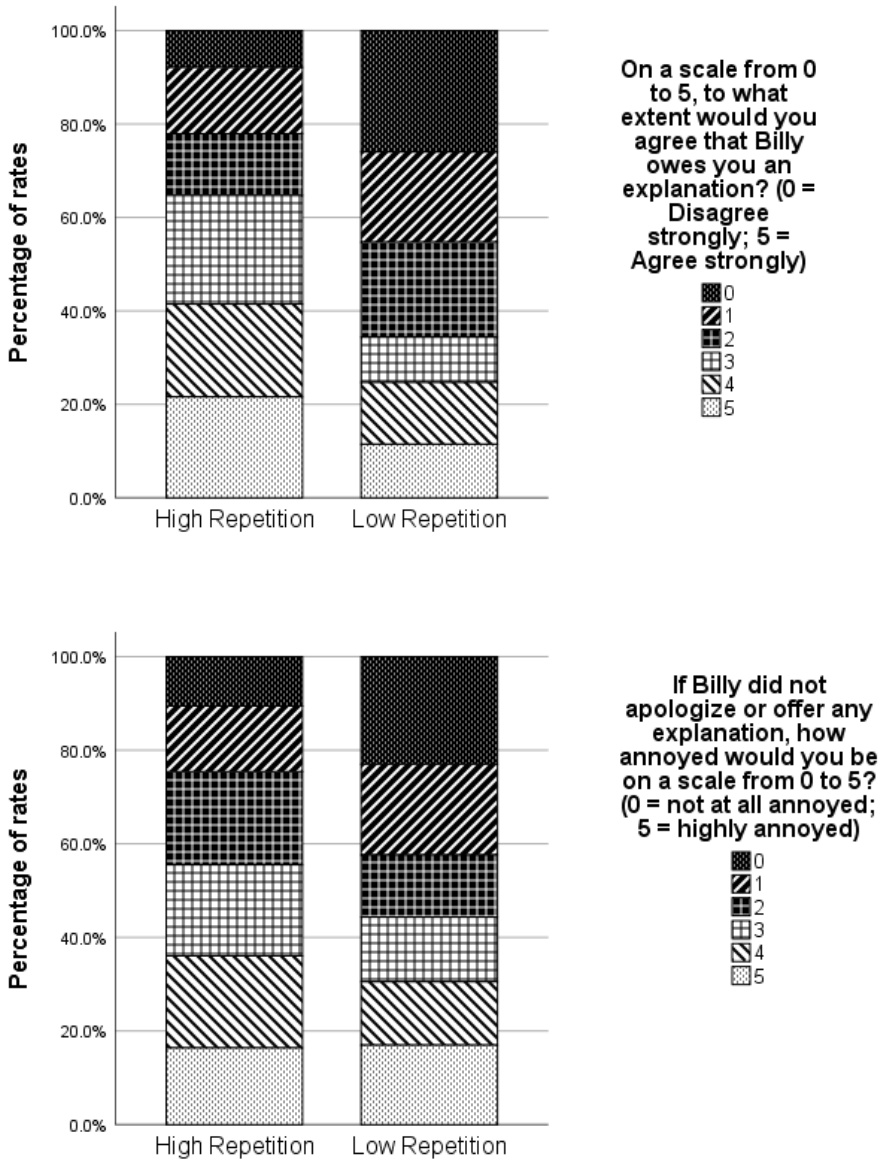

Figure 1.6. Percentage of responses to the normative and affective questions. White background bars indicate a mild-to-strong agreement, whereas black background bars indicate a mild-to-strong disagreement with the statement: in other words, the stronger the agreement, the higher the perception that a commitment has been violated.

As in Study 1c, responses to the implicit question exhibited the opposite pattern to what we had predicted. Participants reported being more willing to restore the previous routine after a commitment had been violated following a longer repeated interaction, giving higher estimates in the High repetition condition (M = 4.04, SD = .96, Mdn = 4) than in the Low repetition condition (M = 3.45, SD = 1.10, Mdn = 4), $t(189) = 4.020$, $p < .001$, Cohen's d = 0.58 (medium effect size). The sample failed the Levene's Test for equality of variance, p = .008. Nonetheless, this pattern of results was confirmed by a nonparametric test, Mann-Whitney U = 3073.000, $p < .001$ (see Figure 1.7).

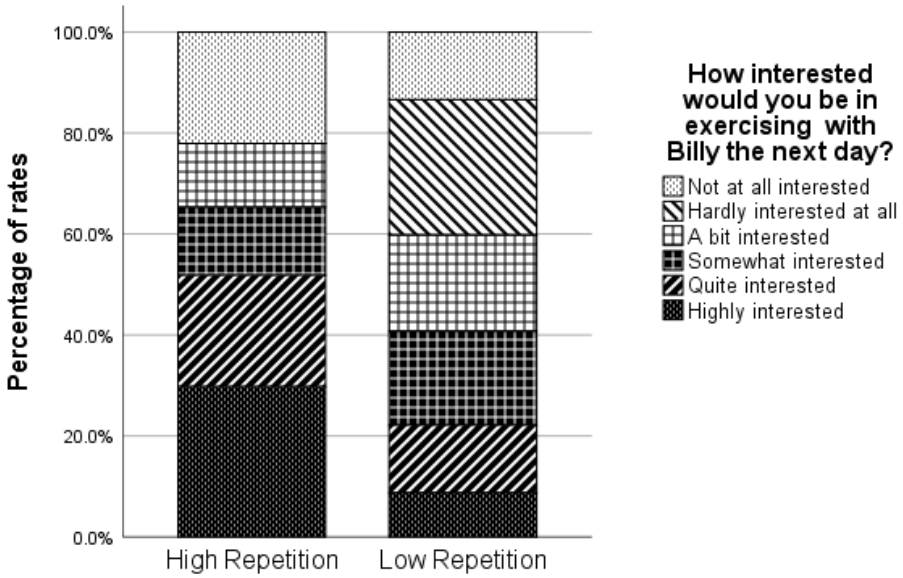

Figure 1.7. Percentage of responses to the implicit question. We reasoned that the stronger the interest in restoring the previous interaction, the greater the perception that a commitment is in place.

As in the previous set of studies, we again found that responses to both the normative question and the affective question tended to cluster around the middle of the scale rather than towards the two extremes (see Figure 1.8). For the normative question, responses tended to be around the midpoint both in the Low Repetition condition (M = 1.53, SD = 1.54, Mdn = 1.00) and in the High Repetition condition (M = 2.31, SD = 1.50, Mdn = 3.00). For the affective question, responses tended to be around the midpoint both in the Low Repetition condition

(*M* = 1.43, *SD* = 1.52, *Mdn* = 1.00) and in the High Repetition condition (*M* = 1.97, *SD* = 1.45, *Mdn* = 2.00).

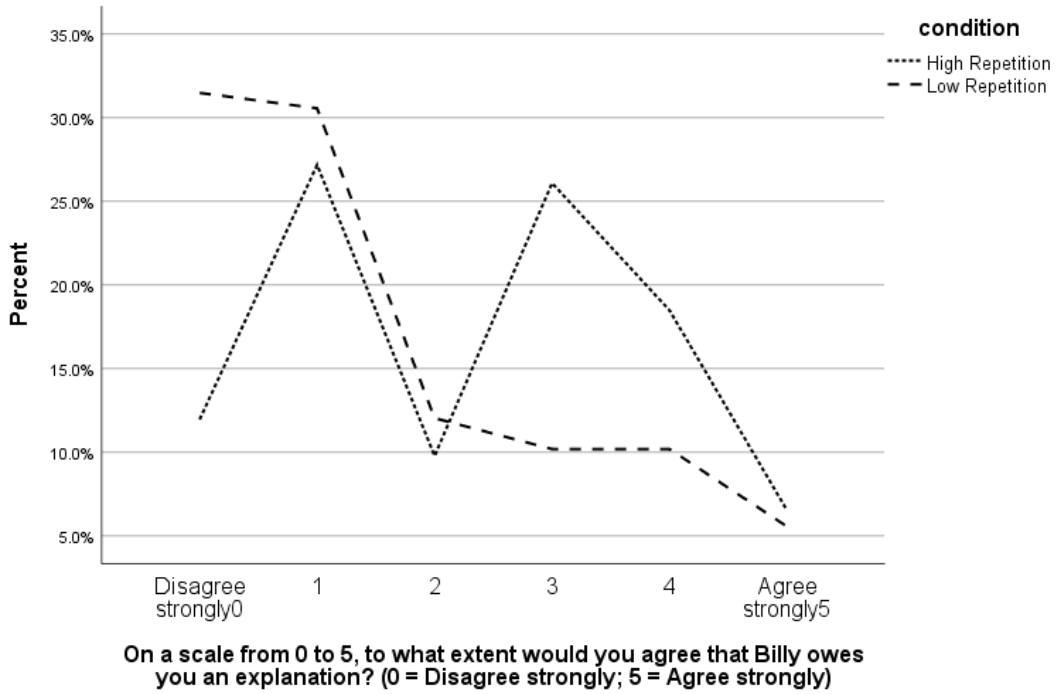

Figure 1.8. Distribution of responses to the normative question. Although in the Low repetition condition there is a significantly higher percentage of responses at the lower end of the scale, we can see that in the High repetition condition the largest number of participants give responses right around the midpoint of the scale.

Discussion

The findings from Study 1d replicate those from Study 1c in a different scenario, which constitutes strong evidence for our hypothesis that commitment attribution can be enhanced as a function of one's history of engagement in a joint activity.

1.5 Discussion of Study 1

In Studies 1a and 1b, we presented participants with vignettes describing a scenario in which one agent had either high expectations (generated by the investment of either a high degree of effort into a joint activity, i.e., the High cost condition) or low expectations (generated by a low degree of effort into a joint activity, i.e. the Low cost condition), and a second agent failed to remain committed. In line with our predictions, the results revealed that participants made more negative normative judgments and reported more negative emotional attitudes in response to the High cost condition than the Low cost condition. Studies 2a and 2b were

designed to probe participants' normative evaluations and affective attitudes in response to scenarios in which one agent failed to remain engaged to a joint activity toward which her partner had either high expectations (generated by a longer history of repeating the routine, i.e., the High repetition condition) or low expectations (generated by a shorter history of repeating the routine, i.e., the Low repetition condition). Again, the results confirmed our predictions: the scenario described in the High repetition condition elicited more negative normative judgments and emotional responses than the scenario described in the Low repetition condition. Taken together, these results provide support of the hypothesis that people's judgments and attitudes about implicit commitments are modulated by cues to others' expectations.

Previous studies suggest that the opportunity cost paid by a partner incentivises prosociality (Charness & Rabin, 2010), and that a partner's apparent investment of effort costs in a joint task increases the people's persistence, as well as their own effort investment, on the task (Chennells & Michael, 2018; Székely & Michael, 2018). These previous findings regarding the relevance of costs for implicit commitment, however, are also consistent with alternative explanations. The costs invested by a partner to engage in a joint task can also be interpreted as a cue to the value of the task itself, leading to higher persistence in the task. The same is true of another convergent line of research showing that participants with a history of successful coordination tend to behave more cooperatively when facing a social dilemma (Guala & Mittone, 2010; Rusch & Lütge, 2016), although it is tempting to interpret such findings as evidence that repeated coordinated interaction might signal reciprocal expectations, and that people may therefore be sensitive to such cues when reasoning about reciprocal commitments. By using both normative and non-normative measures, we were able to rule out alternative explanations. Specifically, our finding that participants were more likely to judge that an apology was in order in the conditions in which we had induced participants to perceive a higher degree of implicit commitment, a pattern consistent with the non-normative, emotional responses, and that cannot be explained by appealing to an increase in the perceived value of the task. In other words, the fact that responses to the normative and the non-normative questions provided a consistent picture suggests that people were not simply expressing their frustration with the outcome presented in the experiment or their disappointment about having missed out on a valuable activity, but that costs and repetition are two factors that are reliably interpreted as cues to others' expectations. Both the normative and the non-normative (affective) questions reliably elicited higher estimates in the High cost/ High repetition conditions. This clearly supports the hypothesis that these two factors enhance people's commitment in joint activity.

In Studies 1a and 1b, investigating the role of costs, responses to our indirect question manifest the same trend, but the difference did not reach statistical significance. This may be because our measure was too weak to pick up on participants' willingness to pay a cost in order to maintain the commitment, or because it was too unrealistic—participants might have enough familiarity with charging phones to assess that four minutes should be enough to be able to send a message. In Studies 2a and 2b, investigating the role of repetition, the implicit measure yielded the *opposite* results to what we had predicted. Our rationale in formulating that question was that the longer the history of repeated interaction, the greater the disapproval of a violation of the routine. This, we predicted, would lead participants to be less inclined to resume the previous routine following a violation. However, the longer history of interaction may also give rise to a more stable commitment relation between partners, which may continue to bind them even after a minor violation. Thus, although the observed results did not confirm our prediction, we believe that they are indeed highly consistent with our hypothesis.

Our results provide further empirical evidence in support of some influential theories of social norms according to which people ought to fulfil others' preferences when they are *reasonably* expected to do so (Lewis, 1969; Bicchieri, 2006). The notion of reasonable expectation is at the core of Lewis' Presumption Reason: Agent A's expectation that agent B will perform an action X is reasonable if A has well-grounded reasons to believe that A will do X. According to Sugden, this moral principle rests upon features of human psychology that enable a motivation to abide by it, such as an aversion to disappoint others' reasonable expectations (Sugden, 2000). And indeed, it has been found that people exhibit an aversion to disappointing others' expectations when those expectations have been made explicit, but only when these expectations were not unreasonable (Heintz et al., 2015). Our results provide further empirical evidence in support of these theories of social norms, demonstrating that people judge there to be an obligation to fulfil others' reasonable expectations even when these expectations have not been made explicit (but have been implicitly cued).

Our findings also have important implications for theorising about the relationship between implicit and explicit commitments (e.g., promises). According to an influential theory of promises (See Scanlon, 1998), the moral ground for the norm that we ought to keep our promises (and, presumably, explicit commitments in general) is that promises generate expectations (i.e., promising to do X creates in the recipient the expectation that the speaker will do X). As shown by our studies, others' (reasonable) expectations also ground implicit commitments in people's moral judgments. Thus, it might be argued that explicit and implicit

commitments share the same moral ground, i.e., that we ought to act in accordance with others' expectations.

Also related to theoretical research on promises, our findings challenge the idea of promissory commitment as a binary notion, according to which either one is committed (i.e., if all conditions for promising are met), or one is not committed (Searle, 1969). This way of conceptualizing promises leaves little room for the idea that recipients' desires and expectations might modulate the promisor's sense of commitment in a graded manner. Since promise-breaking is a violation of a specific norm (i.e., a violation of the norm that one ought to keep one's promises; see Hume, 1739–1740/2000), one might predict that if there was an expectation that the speaker would perform a certain action, violating a promise to perform that action would always be considered blameworthy (on both normative and affective measures) *independently of the magnitude of the expectation.* In contrast to this, we found that given a 6-point scale, participants' assessments of commitment were distributed at intermediate points along the scale rather than at opposite poles. These results foster the idea that for implicit commitments, people assess accountability in a graded manner. Future studies could investigate the effect of recipient's mental attitudes on normative and emotional measures of commitment violation when the commitment has been created by a promise, which may challenge the philosophical conception of promises as binary sources of commitment.

Finally, our findings open up several new avenues for additional further research on implicit commitment. For instance, they raise the question whether different kinds of costs (time, effort, money, etc) may elicit commitment in different ways, which may be reflected in different reparation strategies or in reactions other than moral disapproval. Moreover, while we focused on those costs agents pay to enter into or to carry out a joint activity, it would be interesting to investigate the effects of costs that agents pay as a consequence of commitment violations. Finally, it would be interesting to investigate whether people's responsiveness to cues such as those implemented in our studies has an impact on subsequent partner choice.

To sum up, our studies shed some light on the way people prioritise and evaluate commitments, showing that people are not only sensitive to others' expectations in judging whether commitments are in place, but that they even "sense" commitments when expectations are only *implicitly* cued (e.g., by the amount of costs that one agent is investing in the interaction and by the history of repeated interactions). This sensibility allows people to act together and respond to each others' expectations even in the absence of explicit

agreements, promises, or contracts, and might even be at the basis of the norms that define these acts.

Chapter 2. Knowledge of partner's reliance enhances the perception of commitment

Imagine that you and your friend Kate are planning to meet at the gym to work out together at 6pm. At 5.30pm you discover that some other friends are meeting at the very same time for drinks, and you would prefer to join them, but you also feel you cannot let your friend Kate down. Indeed, she expects to meet you there. She is counting on you. We are often confronted with such choices in everyday life, and our decisions typically involve the feeling that we are committed. We also often find ourselves in situations like that in which Kate finds herself: expecting, counting on, or relying on someone to do something. Commitments are important in a wide variety of social and non-social contexts: We are committed to our partners, our social groups, our jobs, our individual and our shared goals, our values, and even ourselves. Although there are likely to be many similarities across these situations, the current set of studies is restricted to instances of interpersonal commitment—that is, to those commitments that are made by one individual to another individual (Cf., H. H. Clark, 2006).

In the philosophical literature, commitment is usually treated as a relation among one committed agent, one agent to whom the commitment has been made, and an action which the committed agent is obligated to perform in virtue of having given her assurance to the second agent that she would do so (Michael et al., 2016a; cf., Searle, 1969; Scanlon, 1998). Moreover, commitment is treated in this literature as a binary notion: either the aforementioned conditions have been fulfilled (and there is a commitment) or they have not (and there is no commitment). More recently, in the psychological literature, Michael, Sebanz, & Knoblich (2016a) have proposed to treat commitment as a graded phenomenon: One agent can be more or less motivated to perform an action that a second agent is relying on, and may feel more or less guilty if she does not perform the action. To capture this, they introduce the notion of a "sense of commitment", which admits of degrees. In this chapter, we adopt this non-binary conception, as we are interested in people's psychological attitudes about commitment rather than in commitment in the normative sense.

We present empirical results that show what it takes for people to perceive that a commitment has been made. We thus investigate the social conditions that lead people facing standard situations to perceive that a commitment has been made. The act of promising is the canonical way to generate a commitment, and philosophers have analysed the conditions under which a promise is performed and possesses a normative power that commits a speaker to a certain course of action. Speech act theorists claim that this normative power arises when the speaker performs a commissive speech act, that is, a speech act that indicates the speaker's

intention to incur a moral obligation to perform (or omit) a particular action, or that a convention dictates that the given speech act has been performed in such a way and under such circumstances that such obligations have arisen; for instance, stating "I will do it" or nodding after a request are the kind of speech acts (verbal or not) that in the right circumstances are conventionally interpreted as promises (Austin, 1962; Searle, 1969). This raises the question—which is our focus here—what the *right circumstances* are under which people perceive there to be a commitment even in the absence of a commissive speech act.

Several philosophers have pointed out the role of common or mutual knowledge in making a commitment. For instance, Gilbert (Gilbert, 1990, 2006) provides examples where commitments arise from common knowledge of joint goals in the absence of commissive speech acts. MacCormick and Raz (1972) and Scanlon (1998), however, argue that commitments can be formed with neither conventional norms (as when performing a commissive speech act) nor shared goals. What matters, they say, is that one agent leads another agent to form expectations about her future behaviour and to rely on this behaviour. In the current set of studies, we test whether people perceive that a commitment is in place when reliance is mutually known. Our findings indicate that the accounts offered by MacCormick and Raz, as well as by Scanlon, nicely reflect people's judgments when asked to evaluate ecologically valid scenarios.

There is much debate around the notion of "common knowledge". Schelling (1980) and Lewis (1969) point out that coordination games can be solved by assumptions of recursive common knowledge between agents, and Schiffer (1972) defines common knowledge as a hierarchy of propositions that pose strong inferential demands (I know that you know that I know that you know, etc.). However, many acknowledge that agents cannot entertain infinite recursive epistemic states, and several deflationary accounts provide more plausible psychological implementations, such as the availability of the given information in the common ground (Carpenter & Liebal, 2011; Lewis, 1978; Sperber & Wilson, 1986/1995, Chapter 8; Vanderschraaf & Sillari, 2014). Following these cognitively realistic accounts, we understand mutual knowledge in the minimal sense of availability of the information in the common ground, and not as recursive higher-order knowledge. In our experiments, we describe scenarios in which some degree of mutual knowledge of one's reliance is present, that is, in which the agents at least know that one agent will rely on the other agent's behaviour.

Building upon MacCormick and Raz, Scanlon, and Michael, Sebanz, and Knoblich's (Michael et al., 2016a, 2016b) theories, we hypothesise that people have a sense that an agent—the "sender"—is committed to performing X (to believe that the sender is committed, to attribute

blame and to experience negative emotions if the sender does not perform X), if the following conditions are met: (i) The sender has led a second agent (the recipient) to rely on her to do something, and (ii) this is mutually known by the two agents. We operationalise the notion of reliance as the recipient changing her course of action based on her expectations of X occurring. The phenomenon of reliance is often expressed by the recipient with idiomatic expressions such as "I am relying on you", which make explicit the fact that certain expectations are in place, and that the recipient will act accordingly.

In our studies, we consider instances in which it is mutually known that one action of an agent (the sender) has led a second agent (the recipient) to expect her to perform an action X, independently of whether the sender has verbally acknowledged those expectations. In order to investigate whether a recipient's reliance (when it is mutually known) is one factor determining whether a sense of commitment arises, we thus implemented four studies in which we presented participants with scenarios where a sender fails to do X, and manipulated mutual knowledge and the means by which the recipient's expectations were raised: either via an explicit speech act, or through non-verbal events. On the basis of our hypothesis, we predicted that participants' attitudes about whether a commitment has been violated (and about the extent to which the commitment violation warrants blame and more reputational consequences) would depend on whether the recipient's reliance was mutually known, whereas the means by which mutual knowledge has been created would not significantly impact participants' evaluations.

2.1 Study 2a

The first study we conducted was designed to test the hypothesis that mutual knowledge of reliance is a sufficient condition for triggering commitment. To this end, we presented participants with vignettes describing everyday situations in which a sender failed to fulfil the expectations of the recipient. We measured the perception of a commitment being in place by prompting a normative judgment about the sender's behaviour (normative question), by asking whether the situation triggered a feeling of annoyance (affective question), and by probing to what extent the participant herself would be willing to interact with the sender in the future (partner choice question).

Methods

Participants

We implemented a between-subjects design on an online platform (SurveyMonkey, ***http://www.surveymonkey.com)***. Since previous online studies conducted in our lab indicated that non-paid participants present high rates of incomplete and invalid surveys, we opted for a large sample size. A power analysis using G*Power 3.1 (Faul et al., 2007) indicated that a total sample size of 308 participants would be needed to detect a medium effect size (f = 0.25) with a predicted statistical power of 98% using a one-way ANOVA with alpha at .05. Since we planned to run non-parametric tests, we added 15% to our desired sample (Lehmann, 2006). We anticipated that about 25% of participants would not complete the experiment and answer the control questions correctly. Participants were 536 adults, recruited via social media, e-mail, and word of mouth. Data was discarded from participants who did not complete the survey (N = 118) or failed one or more control questions (N = 49), and also from participants who reported being younger than 18 years old (N = 6). This left a total sample size of 364 participants (173 females; M_{age} = 25.80 years, SD = 6.95)—129 in the No mutual knowledge condition, 128 in the Implicit mutual knowledge condition and 107 in the Explicit mutual knowledge condition. The sample was composed for 53.6% by North Americans, for 29.2 % by Europeans, and the rest 17.2 % by participants from other regions.

Here and elsewhere in this chapter, the methods used were in accordance with the international ethical requirements of psychological research and approved by the EPKEB (United Ethical Review Committee for Research in Psychology) in Hungary. All participants gave their informed consent by ticking a box prior to the experiment.

Materials and procedure

Participants were asked to read different hypothetical situations in which a sender violates a recipient's expectations. They were presented with one scenario, in which the agents' expectations either are or are not mutually known, and in which the sender acknowledges these expectations either verbally or only implicitly.

Participants were randomly assigned to one of three conditions: Explicit mutual knowledge, Implicit mutual knowledge, or No mutual knowledge. In the Implicit mutual knowledge condition, the scenario reads as follows:

> *Beth and Ashley are two friends who are planning to go to the seaside for the weekend. Ashley insists on leaving as early as possible because she would like to reach the beach before noon and have lunch there. She offers to pick*

Beth up at 7 a.m. Beth would rather leave at 9 a.m. and have lunch on the way because she hates waking up early. Each of them keeps insisting on her own preference, and they wind up getting mad at each other. The conversation on Friday night ends with Beth telling Ashley "I will wait for you at 9 a.m.", and Ashley telling Beth "I will pick you up at 7 a.m.!". // The same evening Beth goes out to a pub with another friend, who tells her about a nice bistro on the seaside. She then realizes that it could be nice to leave at 7 a.m. after all, and reach the seaside in order to have lunch at this bistro. She sends a message to inform Ashley that she wants to leave early, as Ashley had suggested, and that she will therefore be waiting for Ashley at 7 a.m. When Beth checks her messaging app, she can see that Ashley read the message a couple of minutes after she (Beth) sent it. // In the morning, Beth wakes up early and is ready to go at 7 a.m. As it happens, when Ashley's alarm rings, she decides to turn it off and sleep a bit longer. Ashley arrives at Beth's place at 9 a.m.

In the Explicit mutual knowledge condition, the vignette differed insofar as Ashley replied with a message saying that she would come at 7am, and in the No mutual knowledge Condition the vignette differed in that Ashley did not receive the message (see ***https://osf.io/gsdzb/***for the full vignettes). After reading one of the vignettes, participants were asked to respond to questions about the moral and cooperative character of the agent who changed her course of action (the sender). We hold the sense of commitment to be on a continuous scale rather than a yes/no phenomenon, so we opted for the use of scales as opposed to binary questions. The questions were the following:

- Control question 1: *"At what time did Ashley tell Beth that she would pick Beth up?" ["At 7 a.m."; "at 9 a.m."; "at 11 a.m."].*
- Control question 2: *"At what time did Beth want Ashley to pick her up before she (Beth) learned about the bistro?" ["At 7 a.m."; "at 9 a.m."; "at 11 a.m."].*
- Normative question: *"How wrongly do you think Ashley behaved?" ["Very wrongly"; "A bit wrongly"; "Not particularly wrongly"; "Not at all wrongly]).*
- Partner choice question: *"If you imagine yourself in Beth's situation, would you feel like going on another trip with Ashley in a couple of weeks?" ["Very much"; "A bit"; "Not particularly"; "Not at all"].*

- Affective question: *"If you imagine yourself in Beth's situation, would you feel frustrated/upset/angry towards Ashley?" ["Very much"; "A bit"; "Not particularly"; "Not at all"].*
- Control question 3: *"On the basis of the information that you have, which of the following statements is the most accurate?" ["Ashley did not receive Beth's message about leaving earlier"; "Ashley responded to Beth's message about leaving earlier"; "According to Beth's messaging application, she read the message but did not respond"].*

The normative question was designed to trigger participants' explicit normative judgments about the sender. We predicted that they would evaluate the sender to having misbehaved more often in the two mutual knowledge conditions than in the No mutual knowledge condition, as participants' judgments as to whether a commitment has been violated would depend on whether the recipient's expectations were mutually known. We further predicted that the explicit versus Implicit mutual knowledge conditions would lead to no significant difference in the answers to the questions, as the means by which mutual knowledge was created should be irrelevant for such judgments.

The purpose of the affective and partner choice questions was to control for any mismatch between normative criteria for commitment and a subtler feeling of commitment or emotional disappointment that is not affected by such considerations, as reported by Michael et al. (2016b). The affective question was designed to tap participants' emotional reactions to the violation described. We predicted that they would indicate a higher level of frustration in the two mutual knowledge conditions than in the No mutual knowledge condition, with the additional prediction that there would be no significant difference between the explicit and Implicit mutual knowledge conditions. We reasoned that the violation of a commitment would lead to a negative emotional reaction, and thus the same factors influencing a normative evaluation of the agent's deed would impact participants' levels of frustration.

The partner choice question was designed to probe whether people might engage in a partner choice strategy following the violation of a commitment. We predicted that they would more likely indicate a lower willingness to interact with the sender in the future in the two mutual knowledge conditions than in the No mutual knowledge condition, with the additional prediction that there would be no significant difference between the explicit and the Implicit mutual knowledge conditions. We reasoned that participants would rather avoid interacting with commitment violators, and that the same factors influencing a normative evaluation of the agent's deed would therefore impact participants' partner choices.

The control questions were designed to check whether the participant had read the story with sufficient care to register the information required in order to answer the target questions. Control question 3 was particularly important insofar as it was devised to probe whether participants had understood the critical manipulation. The control and the target questions were presented to the participants in a randomized order, except for the third control question, which was always presented last, since being forced to make a judgment about the epistemic states of the agents could influence the other judgments. Data from those who failed to answer any of the control questions correctly was discarded from the final sample.

Results

To test these hypotheses, we ran a series of Kruskal-Wallis non-parametric tests, and a series of post-hoc tests. Given that our measures involve ordinal scales, we opted for using appropriate non-parametric rather than metric tests (Liddell & Kruschke, 2018). Here and elsewhere in this set of studies, the analyses were performed using IBM SPSS Statistics for Windows v.25.0.0. In accordance with our predictions, a Kruskal-Wallis H test showed that there was a statistically significant difference in the responses to the normative question, $\chi^2(2) = 108$, $p < .001$, $\eta^2 = 0.29$ (large effect size), with a mean rank rate of 110.43 for the No mutual knowledge condition, a mean rank rate of 212.52 for the Implicit mutual knowledge condition and a mean rank rate of 233.48 for the Explicit mutual knowledge condition. In order to determine which condition(s) were responsible for this difference, we ran a series of post hoc pairwise comparison tests showing that responses were significantly lower in the No mutual knowledge condition than in the Explicit mutual knowledge condition ($p < .001$) and in the Implicit mutual knowledge condition ($p < .001$). However, no significant difference was found between the Implicit mutual knowledge condition and the Explicit mutual knowledge condition ($p = .320$) (see Figure 2.1). Here and elsewhere, significance values have been adjusted by the Bonferroni correction. This confirms the hypothesis that the levels of perceived commitment were higher in conditions in which the expectations were mutually known by the agents.

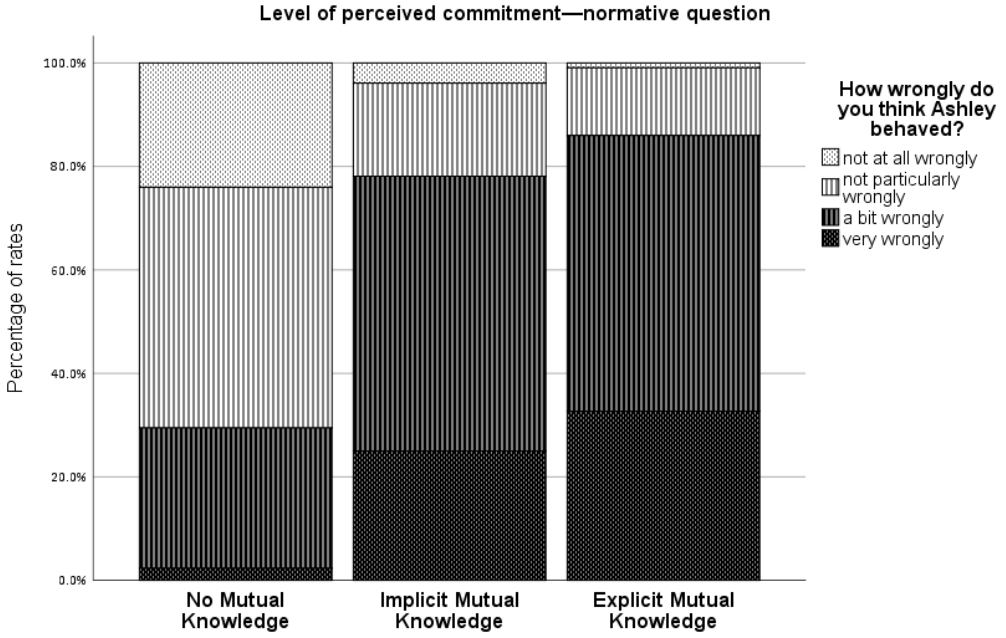

Figure 2.1. Level of perceived commitment—normative question. The responses to the normative question are significantly lower in the No Mutual Knowledge condition than in the Implicit and the Explicit Mutual Knowledge conditions, Kruskal–Wallis Test: $N = 363$, $\chi^2(2) = 108$, $p < .001$, $\eta^2 = 0.29$.

In accordance with our predictions, the responses to the affective question showed the same pattern as for the normative question: A Kruskal-Wallis test revealed that the responses were significantly different in the three conditions, $\chi^2(2) = 83.5$, $p < .001$, $\eta^2 = 0.26$ (large effect size), with a mean rank rate of 119.38 for the No mutual knowledge condition, a mean rank rate of 207.80 for the Implicit mutual knowledge condition and a mean rank rate of 226.75 for the Explicit mutual knowledge condition. Again, a series of post hoc pairwise comparison tests showed that responses are significantly lower in the No mutual knowledge condition than in the Explicit mutual knowledge condition ($p < .001$) and in the Implicit mutual knowledge condition ($p < .001$). No significant difference was found between the Implicit mutual knowledge condition and the Explicit mutual knowledge condition ($p = .773$) (see Figure 2.2). The responses to the affective question predictably correlated with the responses to the normative question, $r_s\,(364) = .53$, $p < .001$.

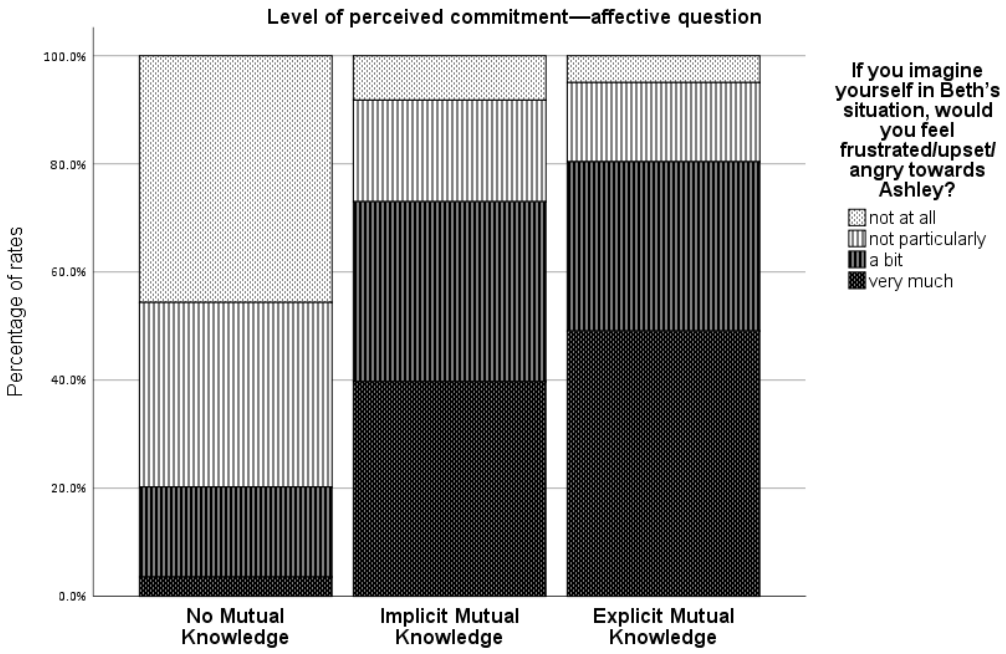

Figure 2.2. Level of perceived commitment—affective question. The responses to the affective question are significantly lower in the No Mutual Knowledge condition than in the Implicit and the Explicit Mutual Knowledge conditions, Kruskal–Wallis Test: $N = 363$, $\chi^2(2) = 83.5$, $p < .001$, $\eta^2 = 0.26$.

The pattern presented above is confirmed for the partner choice question: The responses, tapping participant's willingness to interact again with the sender, were significantly different in the three conditions, Kruskal-Wallis Test, $\chi^2(2) = 40.4$, $p < .001$, $\eta^2 = 0.11$ (medium effect size), with a mean rank rate of 226.01 for the No mutual knowledge condition, a mean rank rate of 152.71 for the Implicit mutual knowledge condition and a mean rank rate of 165.68 for the Explicit mutual knowledge condition. To check that the critical difference lay between the No mutual knowledge condition and the others, we ran a series of post-hoc tests that showed that the responses are significantly higher in the No mutual knowledge condition than in the Explicit mutual knowledge condition ($p < .001$) and in the Implicit mutual knowledge condition ($p < .001$). However, no significant difference was found between the Implicit mutual knowledge condition and the Explicit mutual knowledge condition ($p = .936$) (see Figure 2.3). These results rule out the possibility that participants, while responding to the normative question, were already engaging in some partner choice strategy or implicit disapproval without genuinely evaluating their partner's behaviour as morally wrong.

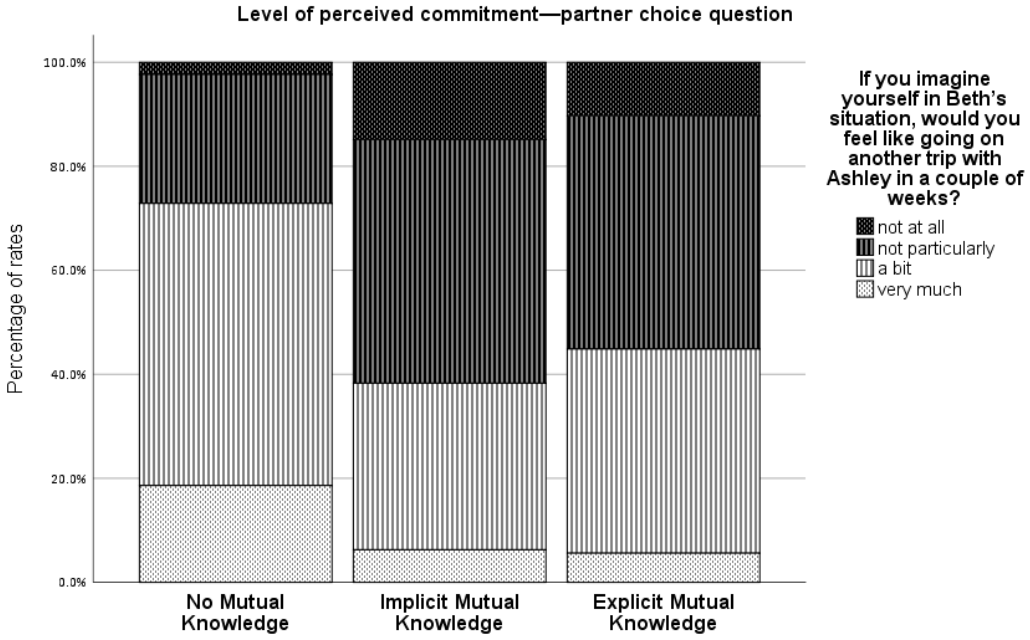

Figure 2.3. Level of perceived commitment—partner choice question. The responses to the partner choice question are significantly higher in the No mutual knowledge condition than in the Implicit and the Explicit mutual knowledge conditions, Kruskal–Wallis Test: $N = 363$, $\chi^2(2) = 40.4$, $p < .001$, $\eta^2 = 0.11$.

The responses to the partner choice question correlated both with the responses to the normative question, $r_s = .422$, $p < .001$, and with the responses to the affective question, $r_s = .459$, $p < .001$.

Discussion

The results corroborated our predictions. Participants evaluated the sender more severely in cases in which the sender had led the recipient to rely on her (and this reliance was mutually known), irrespective of how mutual knowledge had been formed (i.e., whether the sender performed a speech act or not).

Participants' willingness to engage in an unspecified future interaction with the sender was influenced by this factor, but not as strongly as their affective response or their normative evaluation of the sender. There are several possible explanations of this. People take into account several types of information when reasoning about whether one is a desirable partner, information that spans from her competence in a relevant domain (e.g., whether Thomas a good tennis player, if I have to team for a tennis tournament) to her benevolence and willingness to cooperate (e.g., whether Thomas is moved by benevolent intentions) (S. T. Fiske et al., 2007; Heintz et al., 2016). Violating previous commitments is surely among the latter

considerations, but it is reasonable to assume that in the scenario there were other implicit commitments in place between the two agents in addition to the one that was violated, commitments that maybe weight more, as the ones entailed by being friends, and an assumption of reliability due to the (inferred) history of the friendship. When asked about future potential interactions, participants might have taken these factors into account. Furthermore, after responding to the normative question, participants might have been satisfied with having attributed blame to the violator, and therefore considered that an additional precaution would be redundant.

Given that vignettes may be open to a broad range of interpretations, and in light of the inherent noisiness of online data collection, we designed Study 2b to replicate the results of Study 2a using different vignettes.

2.2 Study 2b

Study 2b was designed to implement two different scenarios. Before analysing the data, we ran a preliminary test to check whether the different scenario presented to the participants influenced their responses. An independent-measure Mann-Whitney U test revealed that the responses to the normative question were significantly different between the two scenarios, Mann Whitney: $N = 204$, $U = 2846.5$, $p < .001$. The responses to the affective question were also found to be significantly different, Mann Whitney: $N = 204$, $U = 3539$, $p < .001$, as well as the responses to the partner choice question, Mann Whitney: $N = 204$, $U = 3424$, $p < .001$. These results persuaded us to run additional tests separately and to consider the two scenarios as different studies. We therefore considered the data from the one scenario as Study 2b, and the data from the other scenario as Study 2c.

Compared with Study 2a, in Study 2b we modified an element that might plausibly be relevant to participants' interpretation of the situation, namely the nature of the relationship between the two agents—that is, in Study 2a the two agents were friends, whereas in Study 2b they were colleagues. We implemented mutual knowledge in a similar fashion, that is, via an automatic in-built function of a communication device. This limits the plausible deniability for the sender of not having been exposed to the relevant information.

Methods

Participants

We used SurveyMonkey to implement a web-based paradigm with a between-subjects design. In anticipation of an effect size similar to what was observed in Study 2a, a power analysis using G*Power 3.1 indicated that a total sample size of 231 participants would be needed to detect the expected effect size (f = 0.22) (derived from a predicted statistical power of 85% using a one-way ANOVA with alpha at .05). We added 15% to our desired sample, thus we aimed for a sample size of 265 participants. In total, 265 adults completed the experiment, each of whom was rewarded with $ 0.45. Data was discarded from participants who did not complete the survey (N = 11) or failed one or more control questions (N = 48), and technical errors (N = 2) leaving a total of 204 participants in the final data set. 123 participants were assigned to Study 2 (76 females; M_{age} = 40.67 years, SD = 12.91), 52 in the No mutual knowledge condition, 40 in the Implicit mutual knowledge condition and 31 in the Explicit mutual knowledge condition. As participants were recruited via Amazon M-Turk (***https://www.mturk.com***), the sample was composed entirely by North Americans.

Materials and procedure

As a replication of Study 2a, we followed the very same procedure: Participants were again randomly assigned to one of three between-subjects conditions (Explicit mutual knowledge, Implicit mutual knowledge, No mutual knowledge). In the Implicit mutual knowledge condition, the scenario reads as follows:

> *Betty is a researcher and she is about to attend a workshop in New York along with her team. She is now at the airport, waiting to board her flight. Her colleague and co-presenter Ann will be flying directly to New York from her hometown and meeting Betty and the rest of the team at the workshop. While thinking about her presentation at the boarding gate, Betty realizes that it would be a good idea to include an analysis that Ann did a year earlier. This would help them to impress the team leader at the workshop. // So Betty sends an e-mail to Ann, asking her to bring this material to New York. When Betty arrives in New York, the night before the workshop, she checks her e-mail inbox. She sees that she has received a read receipt from Ann's account, confirming that she (Ann) read the e-mail a couple of minutes after Betty sent it. // As it happens, Ann did not bring her hard-drive with the earlier analysis to New York. So she and Betty do not have this material at*

the workshop, and do not manage to impress their team leader with their results.

In the Explicit mutual knowledge condition, the vignette differs insofar as the sender gives a verbal explicit reassurance to the recipient, whereas in the No mutual knowledge condition the vignette differs as the sender did not receive the information that the recipient was relying on her.

The target questions were the same as in Study 2a, with minor adjustments related to the activity in which the characters were intending to engage. We again controlled for participants' understanding of the text by asking three control questions, the last of which being particularly important because it reveals whether participants understood the critical manipulation.

The questions were presented to participants in a randomised order, except for control question 3, which was always presented last, since we determined that could influence responses to the other questions. Responses from those who failed to answer the control questions correctly were discarded from the final sample.

Results

The results are in line with those of Study 2a. A Kruskal-Wallis Test showed that the responses to the normative question were significantly different in the three conditions, $\chi^2(2) = 25.8$, $p < .001$, $\eta^2 = 0.21$ (medium effect size), with a mean rank rate of 44.36 for the No mutual knowledge condition, a mean rank rate of 71.06 for the Implicit mutual knowledge condition and a mean rank rate of 79.90 for the Explicit mutual knowledge condition. A series of post-hoc pairwise comparisons tests showed that the responses were significantly lower in the No mutual knowledge condition than in the Explicit mutual knowledge condition ($p < .001$) and in the Implicit mutual knowledge condition ($p = .001$). No significant difference was found between the Implicit mutual knowledge condition and the Explicit mutual knowledge condition, ($p = .820$) (see Figure 2.4).

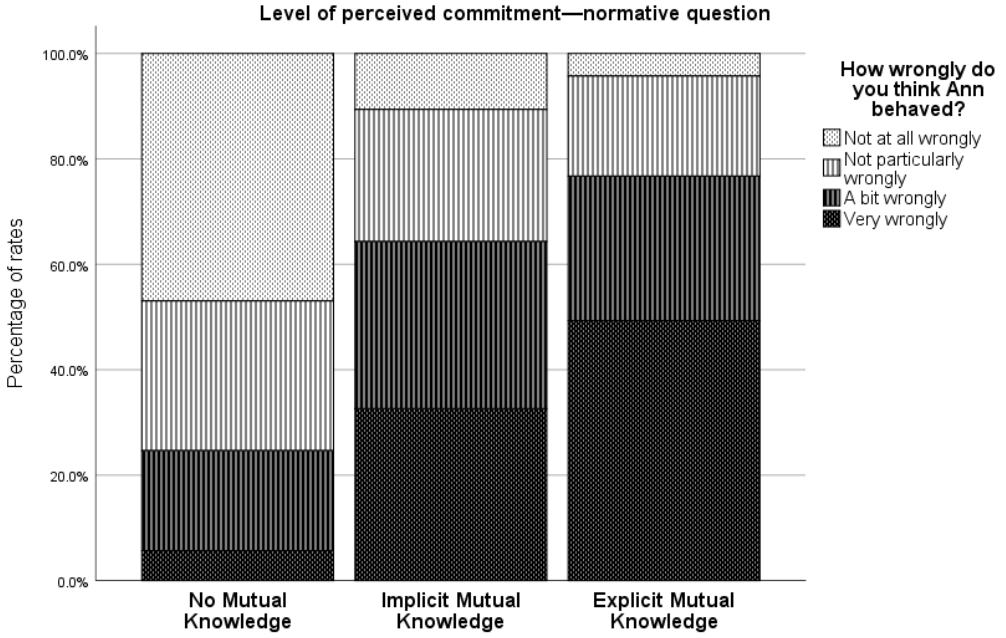

Figure 2.4. Level of perceived commitment—normative question. The responses to the normative question are significantly lower in the No mutual knowledge condition than in the Implicit and the Explicit mutual knowledge conditions, Kruskal–Wallis Test: $N = 123$, $\chi^2(2) = 25.8$, $p < .001$, $\eta^2 = 0.21$.

Consistently with the previous findings, the responses to the affective question showed a similar pattern compared to the normative question: A Kruskal-Wallis Test showed that the responses were significantly different in the three conditions, $\chi^2(2) = 7.5$, $p = .024$, $\eta^2 = 0.06$ (small effect size), with a mean rank rate of 53.30 for the No mutual knowledge condition, a mean rank rate of 64.56 for the Implicit mutual knowledge condition and a mean rank rate of 73.29 for the Explicit mutual knowledge condition. However, a series of post-hoc pairwise comparison tests showed that the responses were significantly lower in the No mutual knowledge condition than in the Explicit mutual knowledge condition ($p = .023$) but not significantly lower than in the Implicit mutual knowledge condition ($p = .315$). as predicted, no significant difference is found between the Implicit mutual knowledge condition (and the Explicit mutual knowledge condition ($p = .808$) (see Figure 2.5). The responses to the affective question predictably correlated with the responses to the normative question, r_s (123) = .584, $p < .001$.

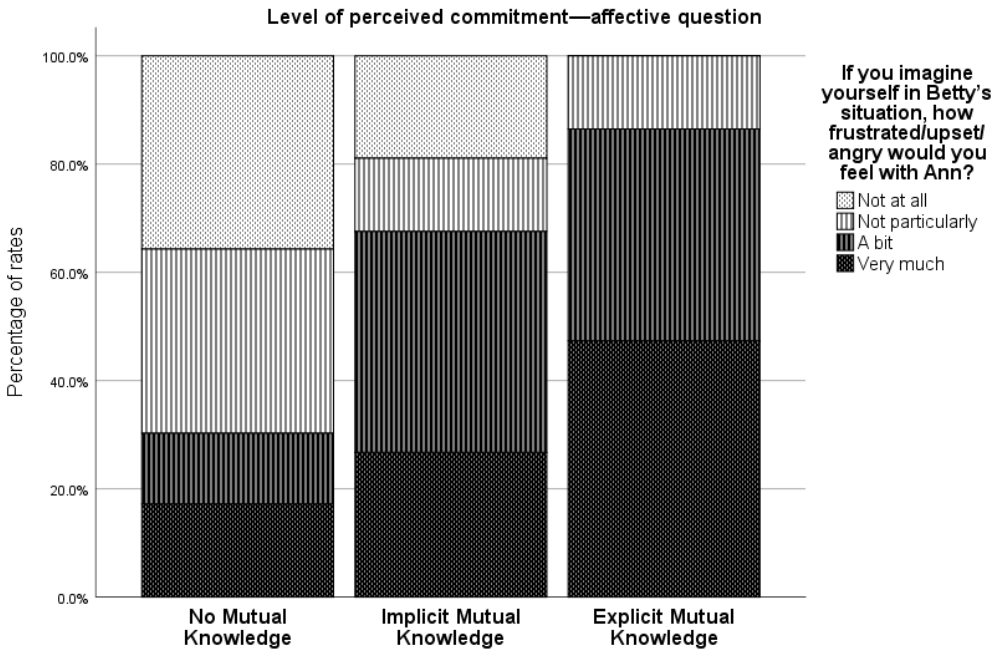

Figure 2.5. Level of perceived commitment—affective question. The responses to the affective question are significantly lower in the No mutual knowledge condition than in the Implicit and the Explicit mutual knowledge conditions, Kruskal–Wallis Test: $N = 123$, $\chi^2(2) = 7.5$, $p = .024$, $\eta^2 = 0.06$.

The responses to the partner choice question confirmed the results found in Study 1: The responses were significantly different in the three conditions, Kruskal-Wallis Test, $\chi^2(2) = 20.3$, $p < .001$, $\eta^2 = 0.17$ (medium effect size), with a mean rank rate of 77.84 for the No mutual knowledge condition, a mean rank rate of 52.15 for the Implicit mutual knowledge condition and a mean rank rate of 48.15 for the Explicit mutual knowledge condition. Again, a series of post hoc pairwise comparisons tests were run. The results showed that the responses were significantly higher in the No mutual knowledge condition than in the Explicit mutual knowledge condition ($p < .001$) and in the Implicit mutual knowledge condition ($p = .001$). Consistently with our hypothesis, no significant difference was found between the Implicit mutual knowledge condition and the Explicit mutual knowledge condition ($p = 1.000$) (see Figure 2.6).

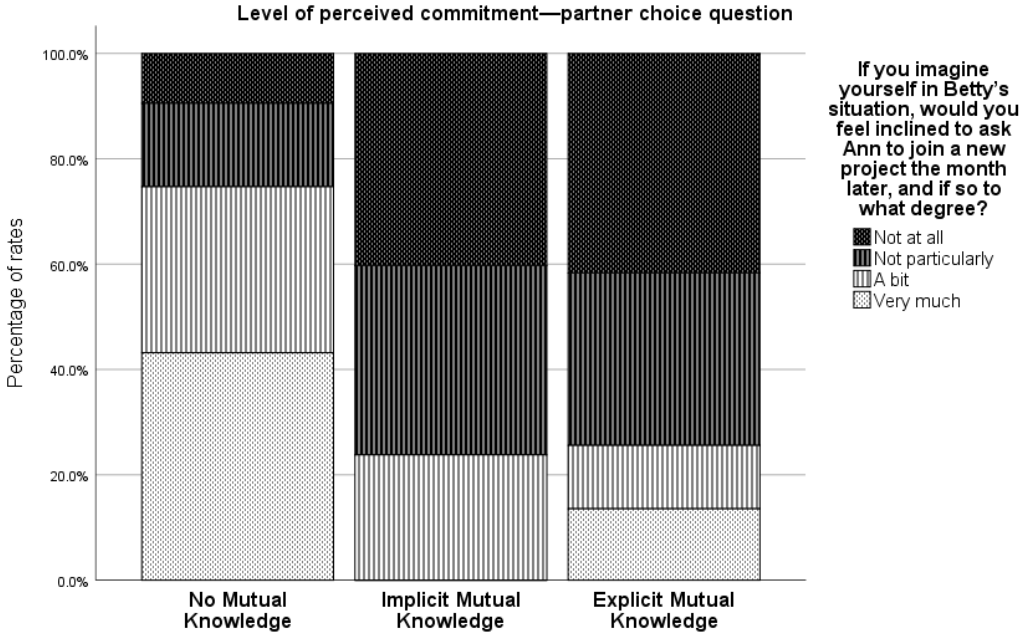

Figure 2.6. Level of perceived commitment—partner choice question. The responses to the partner choice question are significantly higher in the No mutual knowledge condition than in the Implicit mutual knowledge and the Explicit mutual knowledge condition, Kruskal–Wallis Test, $N = 123$, $\chi^2(2) = 20.3$, $p < .001$, $\eta^2 = 0.17$.

The responses to the partner choice question correlated significantly with the responses to the normative question, r_s (123) = .556, $p < .001$, and with the responses to the affective question, r_s (123) = .553, $p < .001$.

Discussion

The results of this second study confirmed our previous findings. The variation in the narrative, as well as the kind of relationship between the two agents, did not affect the pattern we observed previously.

2.3 Study 2c

In Studies 2a and 2b mutual knowledge resulted from a technological device. We designed Study 2c to probe whether commitment can also arise when minimal cues of mutual knowledge are present, such as when it results from a joint attentional process. Participants read descriptions of what we intended to be evidence of mutual knowledge: eye contact, joint attention to a relevant stimulus, and ostensive silence (as suggested by Carpenter & Liebal, 2011). Furthermore, in Studies 2a and 2b the No mutual knowledge conditions present the

following structure: The sender does not lead the recipient to rely on X, and No mutual knowledge about the recipient's reliance is present. To more directly test our claim that a sense of commitment is critically influenced also by the fact that it is mutually known by the agents that the sender had raised the recipient's expectations, Study 2c implemented a situation in which the sender always led the recipient to rely on X: In the No mutual knowledge condition, this is unknown to the sender, while this is mutually known by the agents in both the Implicit mutual knowledge and in the Explicit mutual knowledge conditions.

Methods

Participants

Participants were recruited together with participants for Study 2b. From the original dataset, 81 participants were assigned to Study 2c (44 females; M_{age} = 37.48 years, *SD* = 10.88), 20 in the No mutual knowledge condition, 23 in the Implicit mutual knowledge condition and 38 in the Explicit mutual knowledge condition. As participants were recruited via Amazon M-Turk, the sample was composed entirely by North Americans.

Materials and procedure

In the Implicit mutual knowledge condition, the scenario reads as follows:

> *Jenny and Lisa are two colleagues who work at the same office and get along well. This coming Friday evening, there is an office party taking place in the office lounge. Jenny thinks that it would be a good idea to attend the party, but she usually feels very awkward at such events. Everyone in the office, included Lisa, knows that Jenny always attends parties like this if Lisa, who is very chatty and easygoing, also attends. // On Friday morning, Jenny and Lisa are talking with their boss about the party in the evening. Since Lisa was carrying a couple of bottles of wine to the lounge, Jenny inferred that she was intending to go to the party. So she says to both Lisa and their boss that she will be at the party and that she is looking forward to tasting Lisa's wine. Lisa smiles to her, and the boss replies that he is happy that she (Jenny) will be attending. // However, on Friday afternoon Lisa gets a call from a friend whom she hasn't seen for a long time. Lisa then decides not to go to the party. Jenny is very bored and does not particularly like any of the people at the party. She wishes that she had spent the evening somewhere else.*

The procedure was identical to the one of Study 2b, and the target and control questions were the same as in Study 2a, with minor corrections related to the activity the characters would engage.

Results

The results show very different patterns. A Kruskal-Wallis test revealed that the responses to the normative question were significantly different in the three conditions, $\chi^2(2) = 6.05$, $p = .048$, $\eta^2 = 0.08$ (small effect size). A series of post hoc tests showed no significant differences between each of the three conditions: A marginally significant difference was found between the No mutual knowledge condition and the Explicit mutual knowledge condition ($p = .078$); a non-significant difference between the No mutual knowledge condition and the Implicit mutual knowledge condition ($p = 1.000$); and a non-significant difference between the Implicit mutual knowledge condition and the Explicit mutual knowledge condition ($p = .224$).

In contrast to the previous findings, the responses to the affective question were not significantly different in the three conditions, Kruskal-Wallis test, $\chi^2(2) = 1.325$, $p = .516$. The responses to the affective question correlated significantly with responses to the normative question, $r_s (81) = .573$, $p < .001$.

And again, the responses to the partner choice question were not significantly different in the three conditions, Kruskal-Wallis test, $\chi^2(2) = 3.865$, $p = .145$. The responses to the partner choice question correlated significantly both with the responses to the normative question, $r_s (81) = .281$, $p = .011$, and with the responses to the affective question, $r_s (81) = .358$, $p < .001$.

Discussion

It seems that the changes we implemented in Study 2c influenced participants' responses. The results of Study 2c, which were not predicted, could be explained in three different ways: (a) the way we implemented mutual knowledge may not have been clear to participants—this is partially confirmed by the fact that almost one third of our participants ($N = 32$, 27.6%) failed the comprehension question about the epistemic stance of the sender, thus undermining the reliability of the correct answers; (b) the cues of joint attention we described, that is, eye contact, are not by themselves sufficient cues to mutual knowledge, contrary to previous evidence (Thomas et al., 2014; Siposova et al., 2018); (c) the study on its own lacked the power needed to detect a small effect size; or (d) the fact that the recipient's reliance is the only factor influencing a sense of commitment, provided that these expectations were raised by the sender but irrespective of whether this is mutually known.

We believe that both (a) and (c) are likely explanations. Thus, we ran an additional study to address these concerns. Having only one type of vignette, we maintained a higher sample size to assure that the test would have enough statistical power, and we decided to present the story with a different modality rather than a verbal vignette.

2.4 Study 2d

Given that the inconclusive results of Study 2c might have been due to the way we implemented the manipulation, we decided to replicate the study with a different design. We therefore implemented a different story, in which mutual knowledge was established by a joint attentional process rather than by a technological device. We also chose a different modality rather than a verbal narration of hypothetical events, namely a photo-story, with real people acting out a script. This particular design also has the advantage of increasing the plausibility of the scenario, which is now more likely to be interpreted as something the participants are witnessing rather than merely imagining, thus increasing the ecological validity.

Methods

Participants

We used SurveyMonkey to implement a web-based paradigm with a between-subjects design. In view of the small effect sizes found in the previous studies, a power analysis using G*Power 3.1 indicated that a total sample size of 303 participants would be needed to detect the expected effect size (f = 0.18) (derived from a predicted statistical power of 80% using a one-way ANOVA with alpha at .05). We added 15% to our desired sample, thus we aimed to collect 348 participants. We included data from those participants who had already begun the experiment when M-Turk registered that this number had been reached. Our data set therefore comprised 370 adults, who were rewarded with $0.60 each. Data was discarded from participants who did not complete the survey (N = 15) or who failed one or more control question (N = 117), totalling 238 participants (121 females; M_{age} = 38.30 years, SD = 12.26)—93 in the No mutual knowledge condition, 64 in the Implicit mutual knowledge condition and 81 in the Explicit mutual knowledge condition.

Materials and procedure

Participants were presented with the same basic scenario: For one group of participants the expectations of the agents were not mutually known, for a second group these expectations were mutually known because the sender acknowledged them explicitly, and for

a third group these expectations were mutually known because the sender acknowledged them implicitly.

Participants were randomly assigned to one of the three between-subjects conditions (Explicit mutual knowledge, Implicit mutual knowledge, No mutual knowledge). The scenario was presented as a photo story, as depicted in Figure 2.7.

Figure 2.7. Participants were presented with photo stories which differed according to the three conditions (here an extract from the Implicit mutual knowledge condition).

In the Explicit mutual knowledge condition, the vignette differs insofar as the sender gives an explicit verbal reassurance to the recipient, whereas in the No mutual knowledge condition the vignette differs insofar as the sender was not exposed to the information. The target questions were the same as in Study 2a, with minor adjustments related to the activity the actors were engaged in. We controlled for participants' understanding of the text by asking two control questions. The second control question was particularly important because it revealed whether participants had understood the critical manipulation. Since being forced to make a judgment about the epistemic states of the agents could have an effect on responses to the other test questions, this question was always presented last and on a different page. Except for the second control question, which was always presented last, the questions were presented to the participants in a randomized order. Data from those who failed to answer the control questions correctly were discarded from the final sample.

Results

We predicted that responses to the normative question would be significantly higher in the Explicit mutual knowledge and in Implicit mutual knowledge conditions than in the No mutual knowledge condition. Critically for our hypothesis, we predicted that the rates would not be significantly different between the Explicit mutual knowledge and the Implicit mutual

knowledge conditions. To test these hypotheses, we ran a Kruskal-Wallis non-parametric test and a series of post hoc tests per measure.

Consistently with the predictions, a Kruskal-Wallis test revealed that the responses to the normative question were significantly different in the three conditions, $\chi^2(2) = 34.1$, $p < .001$, $\eta^2 = 0.14$ (medium effect size), with a mean rank rate of 89.49 for the No mutual knowledge condition, a mean rank rate of 139.55 for the Implicit mutual knowledge condition and a mean rank rate of 138.11 for the Explicit mutual knowledge condition). A series of post hoc tests showed that the responses were significantly lower in the No mutual knowledge condition than in the Explicit mutual knowledge condition, $p < .001$); and in the implicit commitment condition ($p < .001$). However, no significant difference was found between the Implicit mutual knowledge condition and the Explicit mutual knowledge condition ($p = 1.000$) (see Figure 2.8). This confirms the hypothesis that the levels of perceived commitment are higher in conditions in which the expectations are mutually known by the agents.

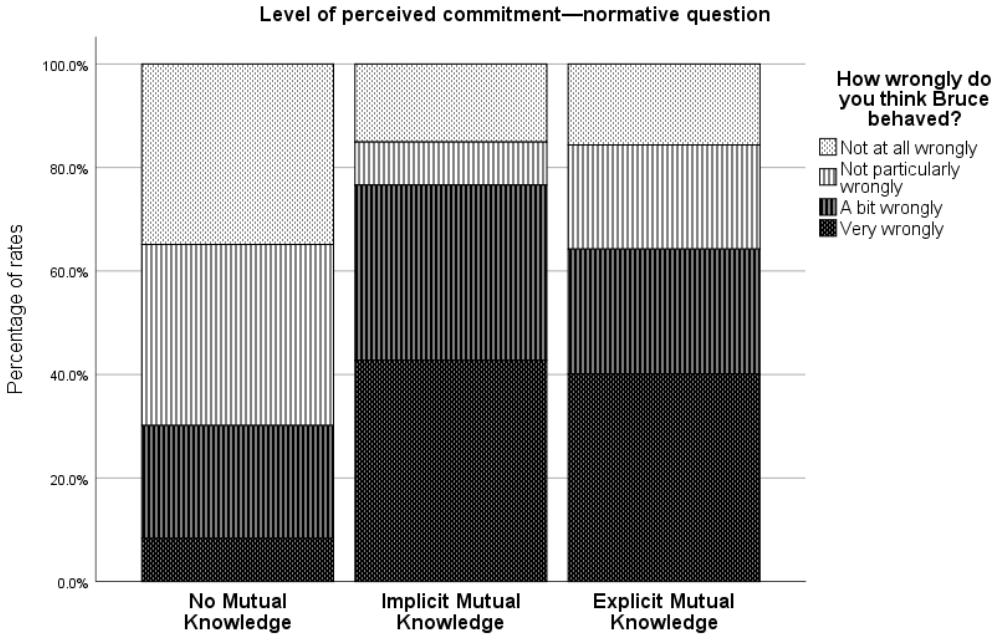

Figure 2.8. Level of perceived commitment—normative question. The responses to the normative question are significantly lower in the No mutual knowledge condition than in the Implicit and the Explicit Mutual Knowledge conditions, Kruskal–Wallis Test: $N = 238$, $\chi^2(2) = 34.1$, $p < .001$, $\eta^2 = 0.14$.

A Kruskal-Wallis test showed that the responses to the affective question were significantly different in the three conditions, $\chi^2(2) = 19.3$, $p < .001$, $\eta^2 = 0.08$ (small effect size), with a mean rank rate of 97.34 for the No mutual knowledge condition, a mean rank rate of

131.95 for the Implicit mutual knowledge condition and a mean rank rate of 135.10 for the Explicit mutual knowledge condition). Consistently with the predictions, the responses showed the same pattern as for the normative question: A series of post hoc tests revealed that responses were significantly lower in the No mutual knowledge condition than in the Explicit mutual knowledge condition (p < .001) and in the Implicit mutual knowledge condition (p = .002). However, no significant difference was found between the Implicit mutual knowledge condition and the Explicit mutual knowledge condition (p = 1.000) (see Figure 2.9). The responses to the affective question were significantly correlated with the responses to the normative question, r_s (238) = .661, p < .001.

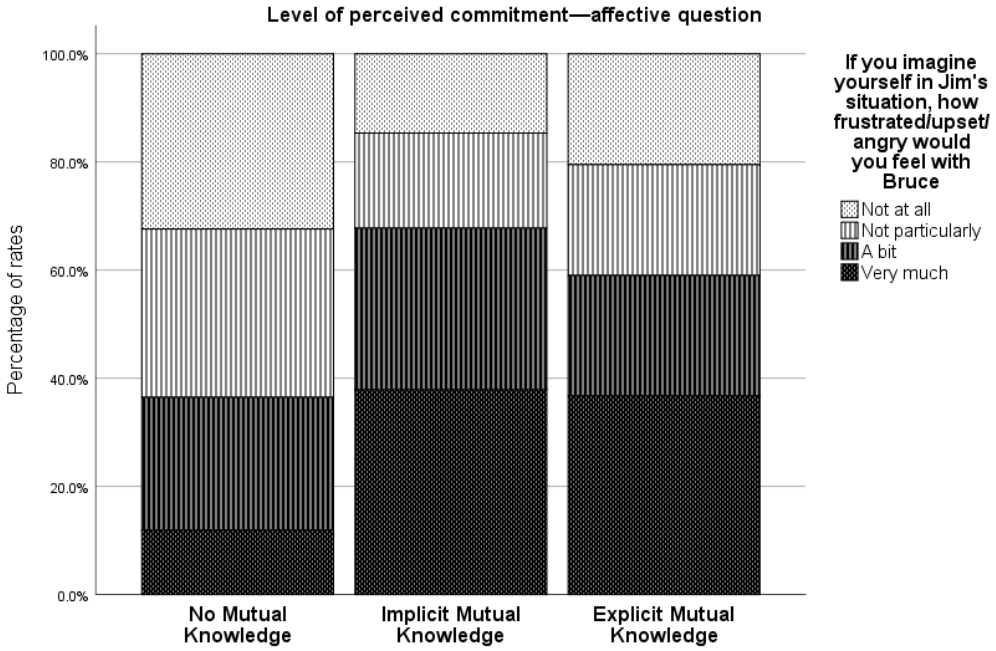

Figure 2.9. Level of perceived commitment--affective question. The responses to the affective question are significantly lower in the No mutual knowledge condition than in the Implicit and the Explicit mutual knowledge conditions, Kruskal–Wallis Test: N = 238, $\chi^2(2)$ = 19.3, p < .001, η^2 = 0.08.

On the other hand, the pattern presented by the partner choice questions was slightly different: The rates of willingness to interact again with the sender were significantly different in the three conditions, Kruskal-Wallis test: N = 238, $\chi^2(2)$ = 9.30, p = .010, η^2 = 0.04 (small effect size), with a mean rank rate of 134.02 for the No mutual knowledge condition, a mean rank rate of 103.98 for the Implicit mutual knowledge condition and a mean rank rate of 115.10 for the Explicit mutual knowledge condition. A series of post hoc pairwise comparisons tests revealed no significant difference between the No mutual knowledge condition and the Explicit

mutual knowledge condition (p = .142), but rates were significantly higher in the No mutual knowledge condition than in the Implicit mutual knowledge condition (p = .010). No significant difference was found between the Implicit mutual knowledge condition and the Explicit mutual knowledge condition (p = .867) (see Figure 2.10).

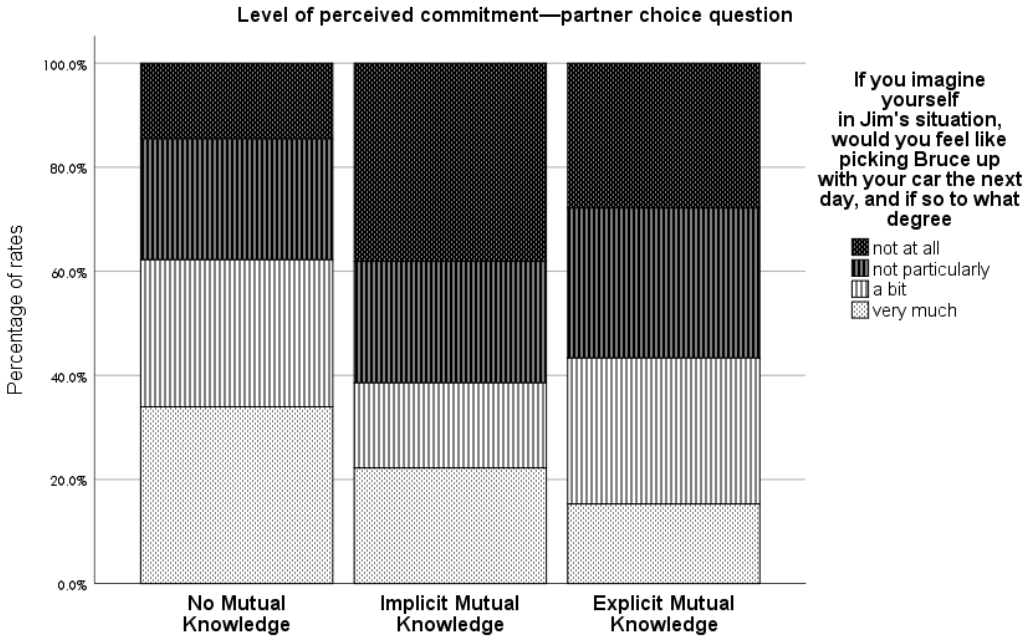

Figure 2.10. Level of perceived commitment—partner choice question. The responses to the partner choice question are significantly higher in the No mutual knowledge condition than in the Implicit and the Explicit mutual knowledge conditions, Kruskal–Wallis Test: N = 238, $\chi^2(2)$ = 9.30, p = .010, η^2 = 0.04.

The responses to the partner choice question were significantly correlated both with the responses to the normative question, r_s (238) = .415, p < .001, and with the responses to the affective question, r_s (238) = .367, p < .001.

Discussion

The results of this study confirm the predictions and replicated the results found in Study 2a and Study 2b. The fact that the recipient's expectations were raised by the sender enhances participant's sense of commitment only when this is mutually known by the agents. The new methodology also conveys the idea that when it is easier to track participants' epistemic states, eye contact is a sufficient trigger of mutual knowledge, as found by Thomas et al. (2014).

2.5 Discussion of Study 2

There are several ways in which a sender can lead a recipient to expect and rely on X, such as uttering a statement that constitutes a commissive speech act, performing an action, or simply omitting to prevent someone from having expectations. For instance, if your friend wants you to attend her party, and you both know that unless you say otherwise, she would expect you to attend, then your silence may be taken to signal your intention to attend, and may thereby generate a sense of commitment to attend. Such cases show that commitment can arise even when the sender does not utter a commissive speech act, such as a promise or an oath, although such acts are efficient means of making expectations mutually known. Likewise, it is not necessary that the sender explicitly acknowledge her recipient's expectations, nor that the sender intended to cause her recipient to expect X for a commitment to arise. For instance, if your dog notices that you are picking up a ball that had been lying on the floor, it is plausible that you will feel committed to playing fetch together, since your action, although unintended, has generated an expectation on the part of your dog that you will play fetch together (see Michael et al., 2016a). Thus, a sense of commitment can arise if the sender leads (voluntarily or not) the recipient to have expectations about her behaviour, if the recipient relies on this expectation, and if this mutually known by them.

This is indeed what we found across a series of four studies. More precisely, we found evidence in support of the hypothesis that the perception of commitment is critically influenced by the extent to which the fact that a recipient has been led by a sender to expect her to do X (Studies 2a and 2b), and that the recipient is going to rely on her to do X, is mutually known (Study 2d). If it is mutually known that a recipient has been led by a sender to expect her to do X, and that the recipient is going to rely on her to doing X, but the sender does not do X, the recipient will hold her accountable. In line with this, the results from our studies indicate that participants evaluated the sender more severely when the recipient's reliance was mutually known than when it was not, irrespective of how their mutual knowledge had been established (i.e., whether or not the sender performed a speech act). It is worth noting that across the four studies the degree of certainty that the agents could have about whether the knowledge was mutual (i.e., whether there was first-order, second-order, or higher-order knowledge about the recipient's reliance) may well have differed—in Studies 2c and 2d, in which mutual knowledge is implemented via cues of joint attention, the degree of certainty is greater than in Studies 2a and 2b, in which mutual knowledge is implemented via in-built features of the technological device used. As much as deniability is reduced in these latter cases, some degree of uncertainty is still present. It is interesting to note, however, that even

in these cases in which it is unclear whether knowledge is mutual or shared, people would often still negotiate in terms of what judgments would be made if knowledge were mutual (see Misyak & Chater, 2014).

Our findings are difficult to reconcile with the hypothesis, suggested by speech act theories, that commitments require speech acts indicating the intention of the speaker to incur a moral obligation to perform a particular action (or to refrain from doing so) (Austin, 1962; Searle, 1969). They are also difficult to reconcile with the conventionalist theories of promises, according to which promising is essentially a socially defined convention enabling coordination and trust within a group (Hume, 1739–1740/2000; Rawls, 1955). While these views differ in important ways, they share at least one important feature—namely, they neglect the phenomenon of unconventional non-verbal commitment.

In contrast, our findings show that Scanlon's account of commitment accurately describes the way people perceive commitment. Scanlon links commitment to the expectations and the reliance of a recipient: According to his theory of promises, the moral norm that we ought to keep our promises is grounded in the fact that promises generate expectations—that is, promising to do something creates in the recipient the expectation that the sender will do it (Scanlon, 1998, pp. 295–302). Our results are also consistent with MacCormick and Raz's claim that when one individual has intentionally led another to rely on her, she is then committed to living up to the other agent's expectation (1972), as well as with Gilbert's analysis, which accords a decisive role to common knowledge in the creation of joint commitments, and which does not require speech acts (1990, 2006). Furthermore, our findings also accommodate some theories of social norms that are grounded in reasonable expectations (Bicchieri, 2006; Sugden, 2000), and they are consistent with previous studies showing that people exhibit an aversion to disappointing others' expectations (Dana et al., 2006; Ockenfels & Werner, 2012), provided that these expectations are not unreasonable (Heintz et al., 2015).

It must be noted that speech act theory, conventionalist accounts of promises and social norm theories are concerned with the normative components of commitment; that is, they do not directly address the issue of its psychological implementation. Thus, none of our findings directly refute these theories. On the other hand, our participants did engage in moral reasoning, which is better captured by an expectation-based explanation. Study 2a and Study 2b implement scenarios in which standardized technology-based signals are used as cues to acknowledge the recipient's expectations. One potential limitation of these scenarios is that these standardized technology-based signals could potentially be interpreted as conventionalized non-verbal speech acts (like nodding), at least in those groups in which they

are commonly used. After all, these signals have the sole function of indicating to the users that the message has been received and read. Study 2d, however, overcomes this limitation, and strengthens the claim that the perception of commitment is not tied to conventional rules or agreements.

Our findings also confirm the prediction that people's assessments of commitment violations influence their partner choices—even in cases in which the commitment was not generated by any speech act. This is important insofar as it highlights the reputational costs of violating commitments even in the absence of speech acts, and thereby also illuminates why agents are so often motivated to honour their commitments (with or without speech acts). In other words, in cases where the expectations of the recipient are mutual knowledge between the recipient and the sender, the sender may anticipate that they would face reputational costs if they did not fulfil these expectations (or at least warn the recipient before disappointing their expectations). Specifically, potential partners in the future may not be willing to rely on them, and may therefore choose not to interact with them. As a result, even in cases in which it would be in the sender's short-term interests not to honour the commitment, the long-term net effects may be negative. This is why mutual knowledge of a recipient's expectations about a sender's future actions—in particular in cases in which the expectations derive from an action performed by the sender—can be sufficient to generate a credible commitment. Interestingly, this point resonates with an observation which Hume made within the framework of contractualism: He noted that when an agent is expected to perform the action that expressed an intention to perform, the agent "subjects himself to the penalty of never being trusted again in case of failure" (Hume, 1739–1740/2000, T 3.2.5.10). While confirming his intuition, our results show that this does not presuppose that commitment-keeping is a conventional practice; it is sufficient if the information flow within the group enables individuals to select their cooperators based on the reputations of the potential partners (Nowak & Sigmund, 2005).

Our findings also open up new avenues for further investigation. Our manipulation was designed to vary whether the recipient's reliance and expectations are mutually known. We implemented mutual knowledge with technology-based signals that limit the plausible deniability of one's knowledge of the partner's reliance, but also with minimal cues of joint attention. Some authors have claimed that eye contact is a potent cue of common knowledge, as eye contact can indicate to both parties that each is aware of the other attending a certain stimulus (in this case, the stimulus is the need of the recipient, and of her reliance on the sender's action) (Siposova et al., 2018; see Carpenter & Liebal, 2011; Thomas et al., 2014). It

would be important for future research to probe the effects of different ways of generating different levels of knowledge.

Moreover, in the scenarios implemented here, the presence or absence of mutual knowledge may also have influenced the degree to which participants attributed expectations to the recipient of the commitment. It is possible, for instance that where expectations are not mutually known, participants may have doubted whether the recipient really expected the sender to perform the action in question. It would be valuable for future studies to manipulate mutual knowledge independently of the strength of expectations.

Part II. Partner's reliance affects the perception of commitment and plausible deniability in communicative contexts

As outlined in the introduction (pp. 1-21) people rely on each other when they perceive other to be committed to do something that is beneficial for those who rely. This perception of others being committed occurs when there is *evidence* of one's incentives to honour their commitment. When people engage in joint activities, this evidence is related to the pursue of the joint goal, namely honouring the commitment translates into doing your part in bringing about the joint goal. The same logic applies to communication, which is itself a joint activity in which partners are providing, receiving or exchanging information.

Communication has been claimed to be a type of joint action or cooperative activity (H. H. Clark, 2006; Grice, 1957; Tomasello, 2008) that follows 'rules' that apply to other cooperative contexts. Both in communicative and cooperative contexts, commitment helps solving the problem of trusting others despite the potential opportunities for others to deceive or defect. Pragmatists and philosophers of language make use of the notion of commitment to refer to the relation that a speaker has with the content conveyed by their communicative act (Austin, 1962; Searle, 1969; Grice, 1957; Brandom, 1994). Performing a communicative act X commits the speaker to certain intentions and beliefs, and to illocutions that derive from X (De Brabanter & Dendale, 2008), and brings about normative effects such as discursive responsibilities and accountability (Geurts, 2019; Marsili, 2016). When performing a communicative act, and thus raising their audience's expectations about the relevance of the communicated content X, a communicator is committed to the relevance of X—relevance that typically overlaps with X being true (Van Der Henst et al., 2002; Wilson & Sperber, 2002).

Perceiving a partner committed to do something that is beneficial, such as providing relevant information, occurs when there is *evidence* of a commitment-reliance relation, such as cues that a partner is relying on the information provided. In the following two chapters I will show how partner's reliance critically affects relevance, and critically affects the perception of commitment; hence the higher the reliance, the higher should be the preventive/compensatory discursive responsibilities for the communicator (see also Geurts, 2019). These preventive responsibilities include for example making certain contextual assumptions more salient to prevent mistaken interpretations to be drawn, using stronger pragmatic cues needed to distance themselves from the message and making amendments if there is evidence that a mistaken interpretation was indeed drawn.

In Chapter 3, I present three experiments which show that participants perceive agent B accountable for a promise violation no matter whether this was explicitly uttered or only implied. Critically, this would not occur when an explicit but non-relied on promise is uttered. In Chapter 4, I present an experiment which shows that strongly implicated promises (e.g., relied on) are perceived as more committal than weakly implicated promises (e.g., less relied on), not only because they bring about a higher degree of accountability for the misleading speaker, but they are as also judged as less plausibly deniable: more relied on implicated promise that are not kept and are further denied cause even higher social consequences for the misleading speaker.

Chapter 3. Speaker commitment to a content is influenced by partner's reliance

Are speakers perceived to be committed only to what they say, or also to what they mean (even when it is not said)? The exact boundaries of the saying-meaning distinction are much discussed in semantics and pragmatics (Austin, 1962; Grice, 1957; Searle, 1969; Sperber & Wilson, 1986/1995; Carston, 2004; Récanati, 2004; Wilson & Sperber, 2004). Some researchers have proposed that this distinction is of key relevance to how commitment is created in communication. In particular, researchers of different backgrounds have proposed that commitments are stronger when meaning is fully linguistically encoded than when it is only implied (Morency et al., 2008; Möschler, 2013; Reboul, 2017). This idea, in short, is that we are more strongly committed to what we say than to what we merely mean. Recent experimental work has been interpreted as providing empirical grounds for this idea (e.g., Lee & Pinker, 2010; Mazzarella et al., 2018). In general, within this literature commitment is understood as speakers' endorsement or distance from what they communicate (see e.g., Boulat & Maillat, 2017).

Here we present theoretical arguments against this picture, and experimental data highlighting clear counter-examples. Specifically, we argue that in the most general perspective what communicators are committed to is the *relevance*[12] of their communicative behaviour, irrespective of whether this is explicitly or implicitly expressed (§3.1). We then present three studies of commitment attribution in the case of promises, contrasting the different roles played by (i) the saying-meaning distinction and (ii) the extent to which an audience relies on what has been expressed (§§3.2-3.5). Participants were presented with vignettes, comic strips and video-clips illustrating everyday situations in which a verbal promise was violated by a communicator. We asked them to judge whether a promise was broken, whether the communicator is a desirable partner for future interaction, and whether the communicator is accountable for any broken promise. We manipulated whether the content of the promise was implicitly or explicitly conveyed and whether the intended audience was likely to rely on the promise. Our findings support the hypothesis that the extent to which the audience relies on the communicator's promise, this being mutually known, is the main factor leading to the attribution of commitment to what has been promised, regardless of whether it is explicit (i.e., linguistically encoded) or implied. This leads us to suggest (§3.6) that the social consequences of the saying-meaning distinction might be overstated. While this distinction is certainly

[12] We use 'relevance' in the sense defined in the Relevance Theory literature, as a trade-off between positive cognitive effects and processing effort (see e.g., Sperber & Wilson, 1986/1995). Truth is a special case of relevance (Wilson & Sperber, 2002).

relevant in some institutionalised contexts, such as rituals and legal texts – where what is explicit often has some privileged status – in most individual social interactions the fundamental factor is instead the relevance of what is expressed (Wilson & Sperber, 2002).

3.1 Commitment and Relevance

The saying-meaning distinction plays a decisive role in many domains of human cultural life. In legal texts and certain ritualised events such as marriage, it is not enough that the agents 'mean' something. Instead, what counts is 'what is said'. One dramatic illustration is the famous Bronston vs. US case (*Bronston v. United States*, 1973). Facing fraud allegations, Mr. Bronston replied to the attorney's question about whether he ever had a personal bank account in Switzerland uttering the following statement: "The company had an account there for about six months, in Zürich". While this statement clearly falsely implied that Bronston never himself had a personal bank account in Switzerland, Bronston was acquitted from a perjury indictment. Strategic uses of the saying-meaning distinction such as this are highlighted by the theory of the strategic speaker, according to which speakers use implicit expressions as a way to provide plausible deniability when it might be advantageous to do so (Lee & Pinker, 2010; Pinker et al., 2008).

Building on these observations, some researchers have argued that communicators are not perceived to be equally committed to what they say and to what they mean. One line of argument is that communicators are perceived as less committed to 'what is meant' because in cases of implicit communication the audience is somewhat responsible for the interpretation of the implicature (Morency et al., 2008). A second line of argument is that since implicatures are cancellable (see Carston, 2004), communicators can use them as pragmatic devices to modulate commitment to the message conveyed (Mazzarella et al., 2018). This in turn creates room for plausible deniability, or provides a means by which the cognitive mechanisms that filter incoming information (called 'epistemic vigilance'; see Sperber et al., 2010) might be bypassed (Reboul, 2017).

There are theoretical reasons to question these arguments. While hearers expect to be provided with relevant information, there is, we suggest, no special reason to expect the information to be fully linguistically encoded (i.e., explicit). It is enough that the information is recognised to be the communicator's intended meaning. This point is most clearly illustrated by cases of non-linguistic communication. If, for instance, you are asked about the quality of a wine and you then raise your eyebrows and produce a sound of satisfaction while tasting it,

you will be perceived as committed to the idea that the wine is good. Such examples speak against the idea that 'what is said' should have any special status for commitment attribution, relative to 'what is meant'. It is true that there are cases in which commitment is tied to 'what is said' rather than 'what is meant', such as those mentioned above (legal texts, marriage rituals, the Bronston case). We suggest, however, that these are special cases in which the means by which commitments take effect is institutionally formalised in one way or another.

There are, moreover, good reasons to positively expect commitments to be tied to 'what is meant', and not 'what is said'. Aspects of the context that are mutually known can influence the interpretation of communicative stimuli, regardless of what is expressed explicitly or implicitly. This can in turn affect commitment attribution. A speaker saying "It's 7.30" when it is mutually known that the audience has to catch a train leaving at 7.32 will (likely) be taken as committed to it indeed being 7.30; but the same utterance in other contexts, including when the departure time is not mutual knowledge, will entail a commitment to it being only approximately 7.30 (and quite possibly later than 7.32). Similarly, when someone makes a promise, they should be perceived to be committed to whatever content makes the promise relevant – which is, largely, the extent that the audience will rely on it being kept. The relevance of what is expressed in a promise depends largely on the extent to which the audience will rely on it in some way, i.e., the extent to which the audience is expected to change their course of action under the assumption that the promise is maintained (the speech act is fulfilled). In other cases—assertions, orders, questions, and so on—the relevance of what is expressed will depend also on other factors, but in all cases what matters for commitment attribution is (we suggest) what is put into the common ground (i.e. 'what is meant') rather than 'what is said'.

All in all, we suggest that in the most general perspective what communicators are committed to what makes their communicative behaviour relevant to the audience (see Van Der Henst et al., 2002; Wilson & Sperber, 2002). Of course, in the case of promises, this commitment is de facto commitment to performing a future action: it is only by acting upon it that the speaker can satisfy their promise (Searle, 1969); this is however not the case for other types of speech act, which have different and more complex relationships between content and action, with varying normative consequences (Bach & Harnish, 1979; but see Geurts, 2019 for a different perspective). Thus, although our study was focused on the specific case of promises, the theoretical point that speakers should be held to be committed to 'what is meant' rather than 'what is said' should generalise to other speech acts too, such as assertions.

3.2 Study 3a

Study 3a tests the hypothesis that the degree to which communicators are taken to be committed to a promise is determined by the degree to which the audience relies on the promise (this being mutually known between the hearer and the speaker), rather than by the degree of explicitness of the promise itself. We contrast promises that are linguistically explicated with promises that need to be pragmatically enriched in order to be understood (more technically, we contrast utterances whose illocutionary force applies to a proposition that does not require any enrichment, with those whose illocutionary force applies to a proposition that does require enrichment). Methodologically, we operationalize commitment by measuring: (1) participants' explicit moral judgements about the communicator, i.e., about whether the communicator should engage in some reparation strategy following the violation of the commitment; (2) whether participants take into account the violation of a commitment when engaging in partner choice, or, in other words, whether the communicator incurred reputational costs; (3) whether participants perceive a violation of a promise. If speakers are committed only to what they say, then these judgements should diverge when 'what is said' differs from 'what is meant'. We test this hypothesis by comparing three conditions (Explicit, Enriched, Explicit but not relied on) which differ in both the degree to which the hearer relies on the promise (i.e., whether they are expected to change their course of action under the assumption that the promise is fulfilled), and whether the promise applies its force to an unenriched proposition or not (i.e., whether or not is fully linguistically encoded).

Methods

Participants

We implemented a web-based paradigm with a between-subjects design on an online platform (SurveyMonkey, ***http://www.surveymonkey.com)***. A power analysis using G*Power 3.1 (Faul et al., 2007) indicated that a total sample size of 279 participants would be needed to detect a low-to-medium effect size (f = 0.2) with a predicted statistical power of 85% using a one-way ANOVA with alpha at .05. Since we planned to run a non-parametric test given the nature of our data (Liddell & Kruschke, 2018), we added 15% to our desired sample (Lehmann, 2006). We thus aimed to collect 321 participants. We included data from all participants who had begun the experiment when we closed the survey collector.

Participants were 322 adults, recruited via SurveyMonkey Audience. Data was discarded from subjects that did not complete the survey (N = 34) and failed one or more control questions (N = 33) totalling 255 subjects (132 females; M_{age} = 49.43 years, SD = 17.58): 81 in

the Explicit condition, 96 in the Enriched condition and 78 in the Explicit but not relied on condition. The sample was composed entirely of North Americans. All participants gave their informed consent by ticking a box prior to the experiment.

The methods used in this and in the following studies are in accordance with the international ethical requirements of psychological research and approved by the EPKEB (United Ethical Review Committee for Research in Psychology) in Hungary.

Materials and procedure

Participants were asked to read different hypothetical situations in which a speaker makes a verbal promise, i.e., performs a commissive speech act, which is later found not to be fulfilled.

Subjects were randomly assigned to one of three between-subjects conditions (Explicit, Enriched, Explicit but not relied on). We manipulated both the degree to which the hearer relies on the promise (i.e., whether they are expected to change their course of action under the assumption that the promise is fulfilled), and whether the promise applies its force to an unenriched proposition or not (i.e., whether or not is fully linguistically encoded).

In the Explicit condition, a speaker makes a promise explicitly, the content of which the hearer will need to rely on. Here, 'what is meant' corresponds to 'what is said', and 'what is said' does not need to be pragmatically enriched. In the Enriched condition, a speaker makes a promise whose content is something the hearer will be likely to rely on. Here 'what is said' needs to be pragmatically enriched in order for the listener to recover 'what is meant'. In the Explicit but not relied on condition, a speaker makes an explicit promise, but the exact content of the promise is not something the hearer will be likely to rely on. Here, 'what is said' is not relevant to attributions of commitment. Then, in both the Explicit and in the Explicit but not relied on condition, the speaker who made the promise fails to comply with the explicitly conveyed content. In the Enriched condition, the speaker who made the promise fails to comply with the conveyed content, but does comply with the explicitly uttered content.

In the Enriched condition, one scenario reads as follows:

> *Jack lent 200 EUR to his friend Ben a couple of months ago. Jack's landlord wants the rent to be paid by Monday of the next week, and Jack does not have enough money to pay it. // So a week before the rent is due, Jack asks Ben to return the money, saying that he wouldn't be able to pay the rent by the Monday deadline otherwise. // Ben replies "Don't worry, I will definitely pay you back." Ben pays back on Thursday, three days after the rent was due.*

In the Explicit condition, the vignette differed only in the addition of words that make the content of the promise explicit—specifically, to "Don't worry, I will definitely pay you back" is added the words "...before Monday". In the Explicit but not relied on condition the vignette was based on the Explicit condition, but differed so that the exact explicit promise will not be relied upon—specifically, the vignette was changed so that the rent is due before Friday, rather than Monday (see ***https://osf.io/bt29w/*** for the full vignettes).

After reading the vignette, participants were given two comprehension questions, and three commitment questions. The comprehension questions were designed to check whether the subject was careful enough in reading the story to have grasped the most relevant information in order to properly respond to the target questions. The commitment questions were designed to measure whether the speaker was perceived to have broken a commitment or not. Specifically, the three questions measure: (i) the extent to which the participant believes that the speaker owes the hearer an apology (Apology Required); (ii) the extent to which the participant would, if in the hearer's position, rely on the speaker in the future (Partner Choice); and (iii) the extent to which the participant believes that the speaker failed to live up to the promise (Violated Promise). For the commitment questions participants indicated, on a 6-point Likert scale, their agreement with some statements about the evaluation of both the situation and the speaker. The responses have been collected on a 6-point scale from 0 (strongly disagree with the statement) to 5 (strongly agree with the statement):

- Comprehension Question 1: *"How much money did Ben borrow from Jack?" ("50 EUR"*; *"200 EUR"*; *"400 EUR")* or *"How long has Alexis been on holiday?" ("One week"*; *"Two weeks"*; *"Three weeks")*
- Comprehension Question 2: *"When did Ben pay the money back ("Monday"*; *"Thursday"*; *"Saturday")* or *"How many times did Bonnie enter the house?" ("Never"*; *"Once"*; *"Four times")*
- Apology Required: "*The speaker owes the hearer an apology*"
- Partner Choice: "*If you were the hearer, you would rely on the speaker in the future"*
- Violated Promise: "*The speaker failed to live up to their promise*"

The comprehension questions and the commitment questions were presented as blocks, in random order. Data from those who failed to answer these questions correctly ($N = 33$) were discarded from the final sample.

For Apology Required, we predicted that participants would more likely evaluate the speaker as having misbehaved in the Explicit and in the Enriched conditions than in the Explicit but not relied on condition, with the additional prediction that there would be no significant difference between the Explicit and the Enriched conditions. For Partner Choice we predicted

that participants would more likely exhibit a lower willingness to interact with the speaker in the future in the Explicit and in the Enriched conditions than in the Explicit but not relied on condition, with the additional prediction that there would be no significant difference between the Explicit condition and the Enriched condition. For Violated Promise, considering the possibility that people might have a naïve intuition of a promise being verbal, we were agnostic about possible differences between the Explicit condition and the Enriched condition. We predicted, though, that participants would perceive a promise being broken significantly more often in the Explicit condition than in the Explicit but not relied on condition, despite the fact that the two statements are identical and both falsified.

Results

To test this hypothesis, we ran a Kruskal-Wallis non-parametric test for each variable; additionally, we ran a Mann-Whitney test to check the effect of scenario for each variable. Given that our measures involve ordinal scales, we opted for using appropriate non-parametric tests (Liddell & Kruschke, 2018). All the analyses from this set of studies were performed using R version 3.4.1 (R Core Team, 2020).

For Apology Required there is a significant difference in the judgements across the two scenarios [Mann-Whitney W (255) = 6759, p = 0.014], with significantly more frequent higher rates in Scenario A than in Scenario B. Thus, we ran the analyses for Apology Required on the two groups separately.

In the Scenario A group, a Kruskal-Wallis H test showed a significant difference in the responses to Apology Required [χ^2 (2) = 15.707, p < .001, η^2 = 0.13]. A series of post-hoc pairwise comparison tests showed that the rates of agreement are significantly lower (i.e., speaker is less frequently judged to owe the hearer an apology) in the Explicit but not relied on condition than in the Explicit condition (p < .001). However, no significant difference is found between the Enriched condition and the Explicit condition (p = 0.09), and between the Explicit but not relied on condition and the Enriched condition (p = 0.09). Here and elsewhere, all p-values were adjusted with the Bonferroni correction for performing three pairwise comparisons. Raw data are illustrated in Figure 3.1.

In the Scenario B group, and consistent with the predictions, a Kruskal-Wallis H test showed that there is a statistically significant difference in the responses to Apology Required [χ^2 (2) = 44.102, p < .001, η^2 = 0.33]. A series of post-hoc pairwise comparison tests showed

that the rates of agreement are significantly lower (i.e., speaker is less frequently judged to owe the hearer an apology) in the Explicit but not relied on condition than in both the Explicit condition (p < .001), and the Enriched condition (p < .001). However, no significant difference is found between the Enriched condition and the Explicit condition (p = 1.00). These results are consistent with our predictions. Raw data are illustrated in Figure 3.1.

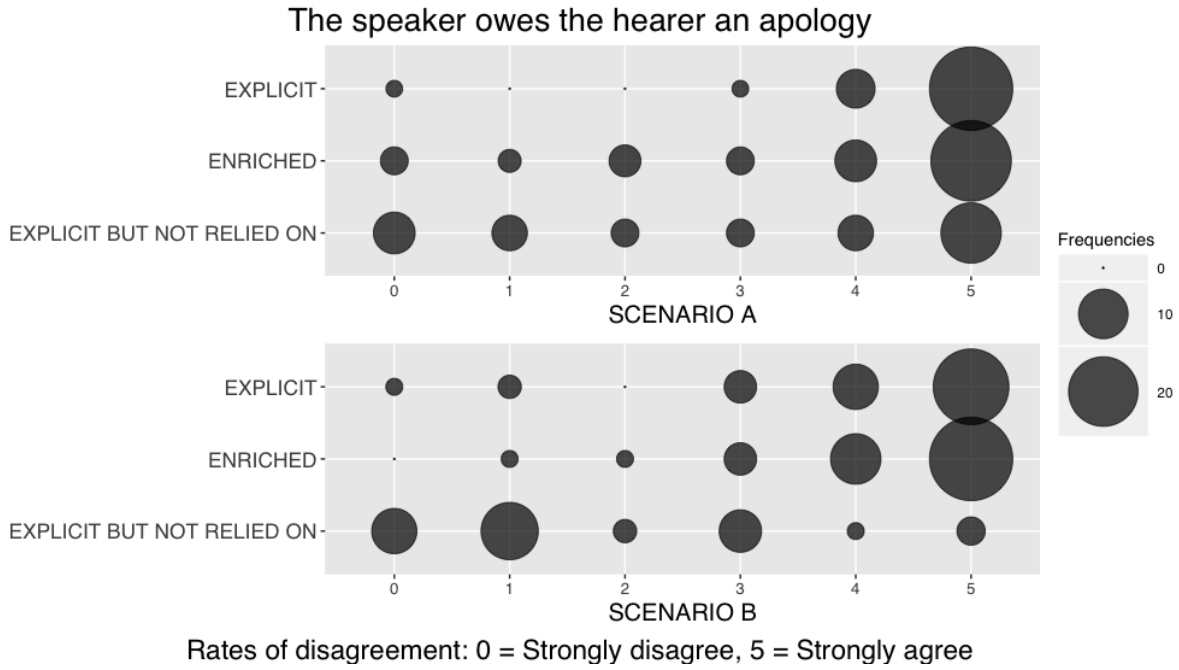

Figure 3.1. Frequencies of responses for Apology Required in the two scenarios. In Scenario A (N = 121), participants were more likely to judge that an apology was in order in the Explicit condition than in the Explicit But not relied on condition, whereas no difference is found between the Explicit condition and the Implicit condition, and between the Implicit condition and the Explicit but not relied on condition. In Scenario B (N = 134), participants were more likely to judge that an apology was in order in both the Explicit condition and the Implicit condition than in the Explicit But not relied on condition, whereas no difference is found between the Explicit condition and the Implicit condition.

The effect of scenario on Partner Choice is non-significant, i.e., we found no significant difference in the judgements across the two scenarios, [Mann-Whitney W (255) = 7493.5, p = 0.272].

Consistent with the predictions, a Kruskal-Wallis H test showed that there is a statistically significant difference in the responses to Partner Choice Question [χ^2 (2) = 48.742, p < .001, η^2 = 0.19]. A series of post-hoc pairwise comparison tests showed that the rates of agreement are significantly higher (i.e., speaker is judged to be significantly more reliable) in the Explicit but not relied on condition than in both the Explicit condition (p < .001) and the Enriched condition (p < .001). However, no significant difference is found between the Enriched

condition and the Explicit condition (p = 1.000) (as shown in Figure 3.2). Raw data are illustrated in Figure 2.

The effect of scenario on Violated Promise is non-significant, i.e., we found no significant difference in the judgements across the two scenarios [Mann-Whitney W (255) = 8422.5, p = 0.558].

A Kruskal-Wallis H test showed that there is a statistically significant difference also in the responses to Perceived Promise [χ^2 (2) = 40.274, p < .001, η^2 = 0.16]. A series of post-hoc pairwise comparison tests showed that the rates of agreement are significantly lower (i.e., the promise is less frequently judged to be broken) in the Explicit but not relied on condition than in both the Explicit condition (p < .001), and the Enriched condition (p < .001). Furthermore, no significant difference is found between the Implicit condition and the Explicit condition (p = .053). Raw data are illustrated in Figure 3.2.

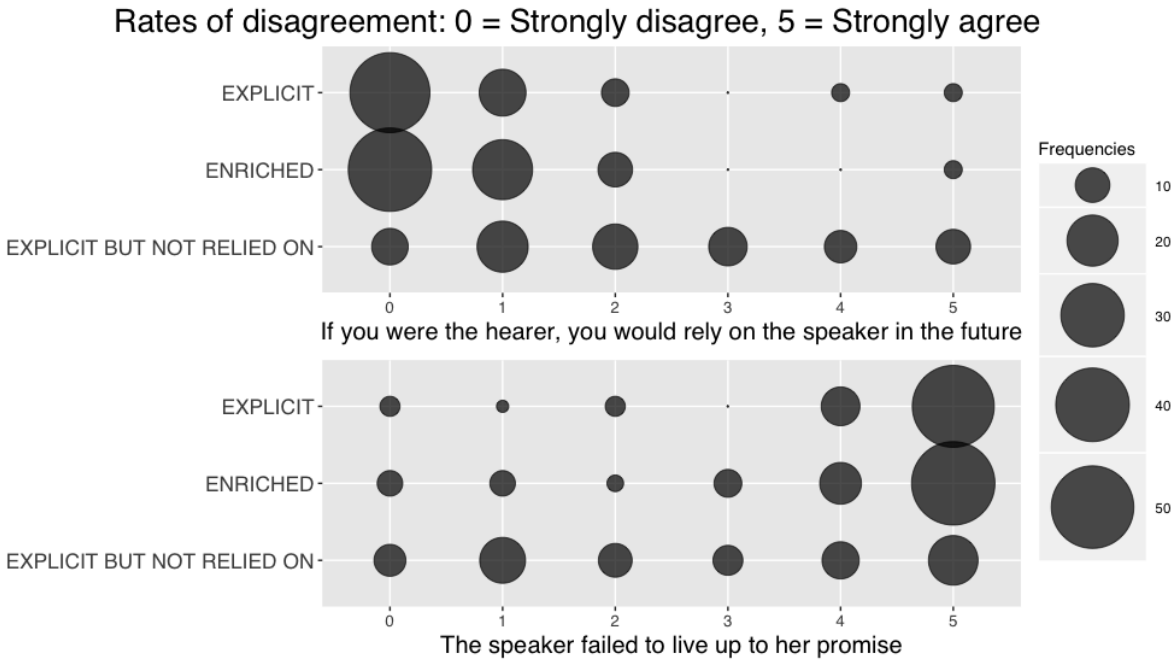

Figure 3.2. Distribution of responses (N = 255) for Partner Choice and Perceived Promise. Participants were less willing to rely on the speaker and more likely to judge that promise was not lived up to both in the Explicit condition and the Implicit condition than in the Explicit But not relied on condition, whereas no difference is found between the Explicit condition and the Implicit condition.

Discussion

Collectively these results support the hypothesis that people take into account pragmatically enriched content when they interpret what the speaker is committing to, and that commitment attribution is modulated by the degree to which the hearer will actually rely on the promise, and not by the degree of its explicitness. However, these results might not hold for all types of pragmatically derived contents. Grice (1989) distinguished Particularised Conversational Implicatures (PCIs) and Generalised Conversational Implicatures (GCIs). PCIs do not contribute to the truth-conditional content of the utterance, typically the explicitly expressed proposition. For instance, in some particular contexts the utterance 'The cake looks delicious' would raise the implicature 'I would like a slice of cake'. GCIs, however, contribute to individuating the truth-condition of a proposition, and therefore some kinds of supposedly implicit content are *de facto* part of 'what is said'. For instance, the utterance 'I ate some of the cookies' generally raise the implicature 'I ate some but not all the cookies'. Furthermore, some studies suggest that people distinguish implicatures (PCIs) from other types of pragmatic operations aimed to retrieve the truth-evaluable proposition (GCIs) (e.g., Doran et al., 2012).

In the next study, we assess whether speakers are perceived to be committed to PCIs, and not just to GCIs. To do this we modify the kind of implicit content used by the characters, contrasting cases in which the promise is fully linguistically available with cases in which the promise needs to be fully retrieved by the hearer.

3.3 Study 3b

Study 3b tests the hypothesis that the degree to which speakers are taken to be committed to a promise is determined not by whether the promise was explicit or not, but by the degree to which the audience relies on the promise. Study 3b thus repeats the general design of Study 3a, but contrasts promises made implicitly with those that are explicit.

Methods

Participants

We implemented a web-based paradigm with a between-subjects design on an online platform (SurveyMonkey). Based on the effect sizes detected in Study 1, a power analysis using G*Power 3.1 indicated that a total sample size of 270 participants would be needed to detect the expected effect size (f = 0.19) (derived from a predicted statistical power of 80% using a one-way ANOVA with alpha at .05). We added 15% to our desired sample, thus we aimed to

collect 310 participants. We included data from all participants who had begun the experiment when SurveyMonkey registered that this number had been reached.

Participants were 310 adults, recruited via Amazon M-Turk (***https://www.mturk.com***). Data was discarded from subjects that did not complete the survey (*N* = 7) and failed one or more control questions (*N* = 7), and technical errors (*N* = 2) totalling 294 subjects (131 females; M_{age} = 37.02 years, *SD* = 10.81): 94 in the Explicit condition, 97 in the Implicit condition and 103 in the Explicit but not relied on condition. The sample was composed entirely of North Americans. All participants gave their informed consent by ticking a box prior to the experiment.

Materials and procedure

Participants were asked to read different hypothetical situations in which a speaker performs a verbal promise, i.e., a commissive speech act, which is later found not to be fulfilled. Subjects were presented with one of the two scenarios presented below, in one of three conditions (Explicit, Implicit, Explicit but not relied on) in which the promise is explicit rather than only implied, and in which the relevance of the promise is high rather than low.

Subjects were randomly assigned to one of three between-subjects conditions (Explicit, Implicit, Explicit but not relied on).

In the Implicit condition, one of the scenarios reads as follows:

> *Andrea is working on her Master thesis and the final draft is almost done. Because Andrea is not a native English speaker, she asks another student, Jen, to proofread the draft. She asks: "Can you help me out and check my writing? I'll have to hand in my thesis in three days." Jen answers: "I have some free time tomorrow". Andrea receives the proof-read document from Jen four days later, one day after her deadline.*

In the Explicit condition, the vignette differed only to the extent that the implied information was now explicitly uttered: "I'll bring some food" ("I'll read it then"). In the Explicit but not relied on condition the vignette was based on the Explicit condition, but differed so that the explicit promise will not be relied upon -- specifically, the vignette was changed so that satisfying the statement "I'll bring some food" ("I'll read it then") would be irrelevant for the hearer—as in Scenario A the speaker makes clear there will be food for everyone, and in Scenario B the thesis deadline is still two weeks ahead (see ***https://osf.io/zrufe/*** for the full vignettes).

As in Study 3a, after reading the vignette, participants were asked to indicate their agreement with some statements. We measured whether the speaker was perceived to have broken a commitment or not. The responses have been collected on a 6-point Likert scale from 0 (strongly disagree with the statement) to 5 (strongly agree with the statement).
The commitment measures were the same as in Study 3a. We again controlled for participants' understanding of the text by asking two comprehension questions (see ***https://osf.io/zrufe/***). Both the comprehension questions and the commitment measures were always presented in a randomised order. Responses from those who failed to answer this question correctly were discarded from the final sample (N = 7).

Results

To test this hypothesis, we ran a Kruskal-Wallis non-parametric test for each variable; additionally, we ran a Mann-Whitney test to check the effect of scenario for each variable.

The effect of scenario on Apology Required is non-significant, i.e., we found no significant difference in the judgements across the two scenarios [Mann-Whitney W (293) = 10432, p = 0.65]. Consistent with the hypothesis, a Kruskal-Wallis H test showed a significant difference in the responses to Apology Required [χ^2 (2) = 124.85, p < .001, η^2 = 0.43]. A series of post-hoc pairwise comparison tests showed that the rates of agreement are significantly lower (i.e., speaker is less frequently judged to owe the hearer an apology) in the Explicit but not relied on condition than in the Explicit condition (p < .001). However, no significant difference is found between the Implicit condition and the Explicit condition (p = 0.09), and between the Explicit but not relied on condition and the Implicit condition (p = 0.09). Raw data are illustrated in Figure 3.3.

The effect of scenario on Partner Choice is non-significant [Mann-Whitney W (293) = 12066, p = 0.062]. Consistent with the predictions, a Kruskal-Wallis H test showed that there is a statistically significant difference in the responses to Partner Choice Question [χ^2 (2) = 83.896, p < .001, η^2 = 0.29]. A series of post-hoc pairwise comparison tests showed that the rates of agreement are significantly higher (i.e., speaker is judged to be significantly more reliable) in the Explicit but not relied on condition than in both the Explicit condition (p < .001) and the Implicit condition (p < .001). However, no significant difference is found between the Implicit condition and the Explicit condition (p = 1.000) (as shown in Fig. 2). Raw data are illustrated in Figure 3.3.

The effect of scenario on Violated Promise is non-significant [Mann-Whitney W (293) = 12048, p = 0.06]. A Kruskal-Wallis H test showed that there is a statistically significant difference also

in the responses to Perceived Promise [χ^2 (2) = 78.802, p < .001, η^2 = 0.27]. A series of post-hoc pairwise comparison tests showed that the rates of agreement are significantly lower (i.e., the promise is less frequently judged to be broken) in the Explicit but not relied on condition than in both the Explicit condition (p < .001), and the Implicit condition (p < .001). A significant difference is found also between the Implicit condition and the Explicit condition (p = .046). Raw data are illustrated in Figure 3.3.

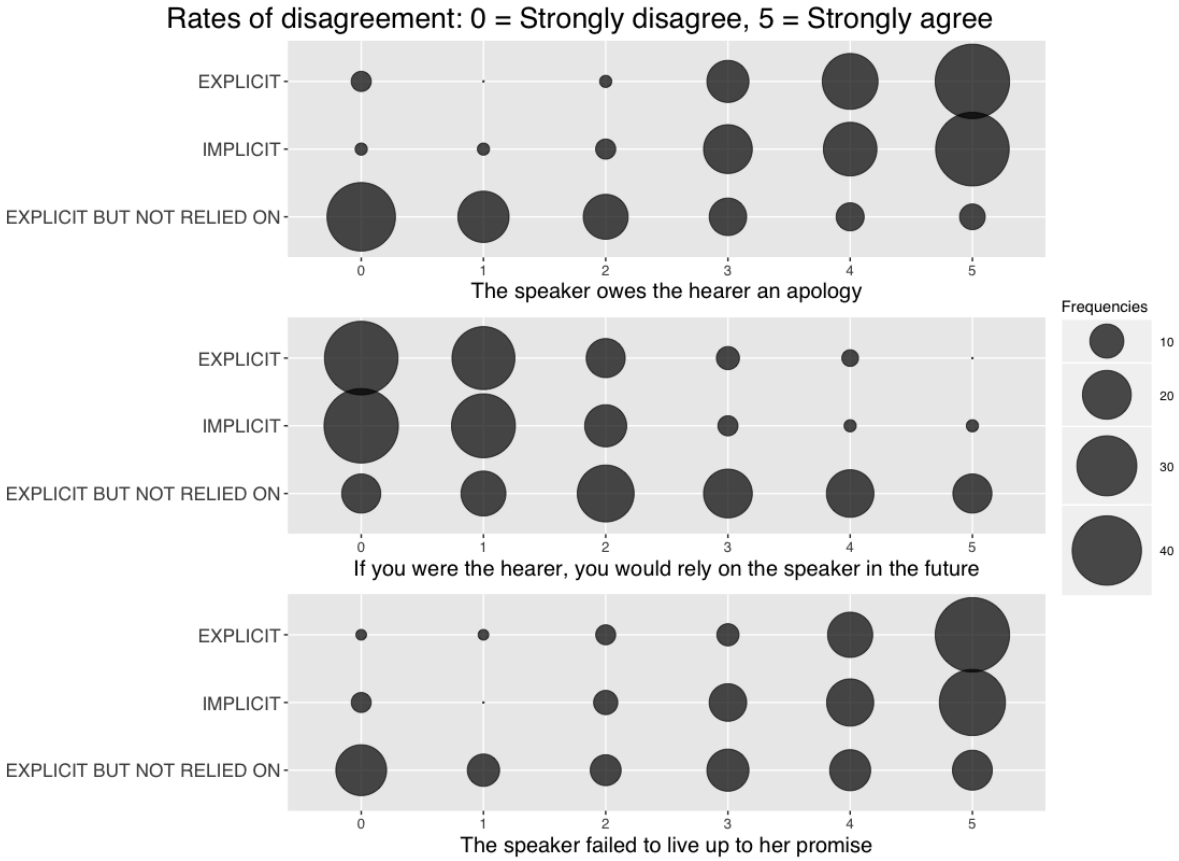

Figure 3.3. Distribution of responses (N = 294) for Study 3b. As can be seen, participants were more likely to judge that an apology was in order and that the speaker was unreliable in the Implicit and in the Explicit conditions than in the Explicit But not relied on condition, whereas no difference is found between the Explicit condition and the Implicit condition. However, participants were more likely to judge that promise was not lived up to in the Explicit condition than in the other two conditions, and more likely to judge the same in the Implicit condition than in the Explicit but not relied on condition.

Discussion

Collectively these results support the hypothesis that with regards to commitments, there is no principled or qualitative distinction between explicit and implicit communication.

The one finding that was not replicated concerns the perception of a violation of a promise; while in Study 3a, the pragmatically enriched content was taken into account in the

assessment of the promise (with no significant differences between the Explicit and the Implicit conditions), in Study 3b the pragmatically inferred content was less taken into account by participants (they were more likely to judge the promise as violated in the Explicit condition than in the Implicit condition).

In order to increase the robustness and generalise our results, we ran another study, effectively a conceptual replication of the previous study, using a different type of stimuli.

3.4 Study 3c

Study 3c was designed to implement two different scenarios. We first checked whether the scenario we presented significantly influenced participants' responses. For all three measures there is a significant difference in the judgements between the two scenarios. There is a significant difference in the responses to Partner Choice across the two scenarios [Mann-Whitney W (310) = 7806, $p < .001$], with significantly higher rates in Comic Scenario than in Video Scenario. There is a significant difference in the responses to Apology Required across the two scenarios [Mann-Whitney W (310) = 16857, $p < .001$], with significantly higher rates in Video Scenario than in Comic Scenario. Finally, there is a significant difference in the responses to Perceived Promise across the two scenarios [Mann-Whitney W (310) = 17376, $p < .001$], with significantly higher rates in Video Scenario than in Comic Scenario.

Since the scenario played a role in the formation of the judgements, we decided to run the analysis for the two scenarios independently despite the loss of statistical power. We therefore considered the data from the first scenario as Study 3c (N = 153), and the data from the other scenario as Study 3d (N = 157).

As did Study 3b, Study 3c also tests the hypothesis that the degree to which speakers are taken to be committed to a promise is determined not by whether the promise was made explicitly or not, but by the degree to which the audience relies on the promise. Study 3c replicated the design of Study 3a and 3b, but differed in the following respects: (i) differently to Study 3a, the kind of implicit content used is a conversational implicature (a PCIs); and (ii) we presented participants with video-clips rather than with written scenarios. This increases the ecological validity of the situation; and underlines the mutual manifestness of the fact that the hearer relies on the promise, by creating a clearly perceptible case of joint attention (Carpenter & Liebal, 2011) rather than relying on explicit verbal description of the agents' epistemic stances.

Methods

Participants

We implemented a web-based paradigm with a between-subjects design on an online platform (SurveyMonkey). Based on the effect sizes detected in Study 3a, a power analysis using G*Power 3.1 indicated that a total sample size of 270 participants would be needed to detect the expected effect size (f = 0.19) (derived from a predicted statistical power of 80% using a one-way ANOVA with alpha at .05). We added 15% to our desired sample, thus we aimed to collect 310 participants. We included data from those participants who had already begun the experiment when SurveyMonkey registered that this number had been reached.

Participants were 346 adults, recruited via Amazon M-Turk. Data was discarded from subjects that did not complete the survey (N = 19) and failed one or more control questions (N = 17) totalling 310 subjects[13]. From the original dataset, 153 participants were assigned to Study 3c, 45 in the Explicit condition, 67 in the Implicit condition and 41 in the Explicit but not relied on condition. The sample was composed entirely of North Americans. All participants gave their informed consent by ticking a box prior to the experiment.

Materials and procedure

Participants were asked to watch different hypothetical situations involving a potential violation of a promise. Subjects were presented with one of the two scenarios described above, in one of three conditions (Explicit, Implicit, Explicit but not relied on) in which the promise is explicit, or only implied, and in which the relevance of the promise is high rather than low.

Subjects were randomly assigned to one of three between-subjects conditions (Explicit, Implicit, Explicit but not relied on). The scenario is presented to be read as if the events occurring in the scenario were true.

In the Implicit condition, the scenario was shown as follows:

> [In a black and white short video-clip, X and Y are having a chat at a café and at some point X mentions that he needs to go to the restroom. He points at his laptop on the table and looks at Y. Y states “You can leave it here”. When X comes back from the restroom, the laptop is on the table unattended while Y is chatting with another person.]

[13] Due to human error, some demographic info was not collected.

In the Explicit condition, the vignette differed only to the extent that the implied information was now explicitly uttered: “I’ll keep an eye on it". In the Explicit but not relied on condition the vignette was based on the Explicit condition, but differed so that the explicit promise will not be relied upon—specifically, the vignette was changed so that satisfying the statement (“I’ll keep an eye on it”) would be irrelevant for the hearer—as there were two other friends keeping an eye on the laptop at the table[14] (see ***https://osf.io/zrufe/*** for the full vignette).

As in Study 3a and 3b, after reading the vignette, participants were asked to indicate their agreement with some statements. We measured whether the speaker was perceived to have broken a commitment or not. The responses have been collected on a 6-point Likert scale from 0 (strongly disagree with the statement) to 5 (strongly agree with the statement). The commitment measures were the same as in Study 3a and 3b. We again controlled for participants’ understanding of the text by asking two comprehension questions (see ***https://osf.io/zrufe/***). Both the comprehension questions and the commitment measures were presented always in a randomised order. Responses from those who failed to answer this question correctly were discarded from the final sample (N = 11).

Results

To test our hypothesis, we ran a Kruskal-Wallis non-parametric test for each variable. Consistent with the predictions, a Kruskal-Wallis H test showed that there is a statistically significant difference in the responses to Apology Required [χ^2 (2) = 6.291, p = .043, η^2 = 0.04]. However, a series of post-hoc pairwise comparison tests showed a non-predicted pattern: the rates of agreement are significantly lower (i.e., speaker is less frequently judged to owe the

[14] However, it has been noticed that X may need assurance that somebody is going to be responsible for minding the laptop. In the Explicit but not relied on condition, two other characters happen to be there, but they have given no assurance about it, making Y’s promise still highly relevant. This interpretation would of course affect the results.

hearer an apology) in the Explicit but not relied on condition than in the Implicit condition (p = .043), but not than in the Explicit condition (p = .257). As predicted, no significant difference is found between the Implicit condition and the Explicit condition (p = 1.000). Raw data are illustrated in Figure 3.4.

Contrary to the predictions, a Kruskal-Wallis H test showed that there is no statistically significant difference in the responses to Partner Choice [χ^2 (2) = .944, p = .624]. A Kruskal-Wallis H test showed that there is no statistically significant difference in the responses to Perceived Promise [χ^2 (2) = 4.155, p = .125]. Raw data are illustrated in Figure 3.4.

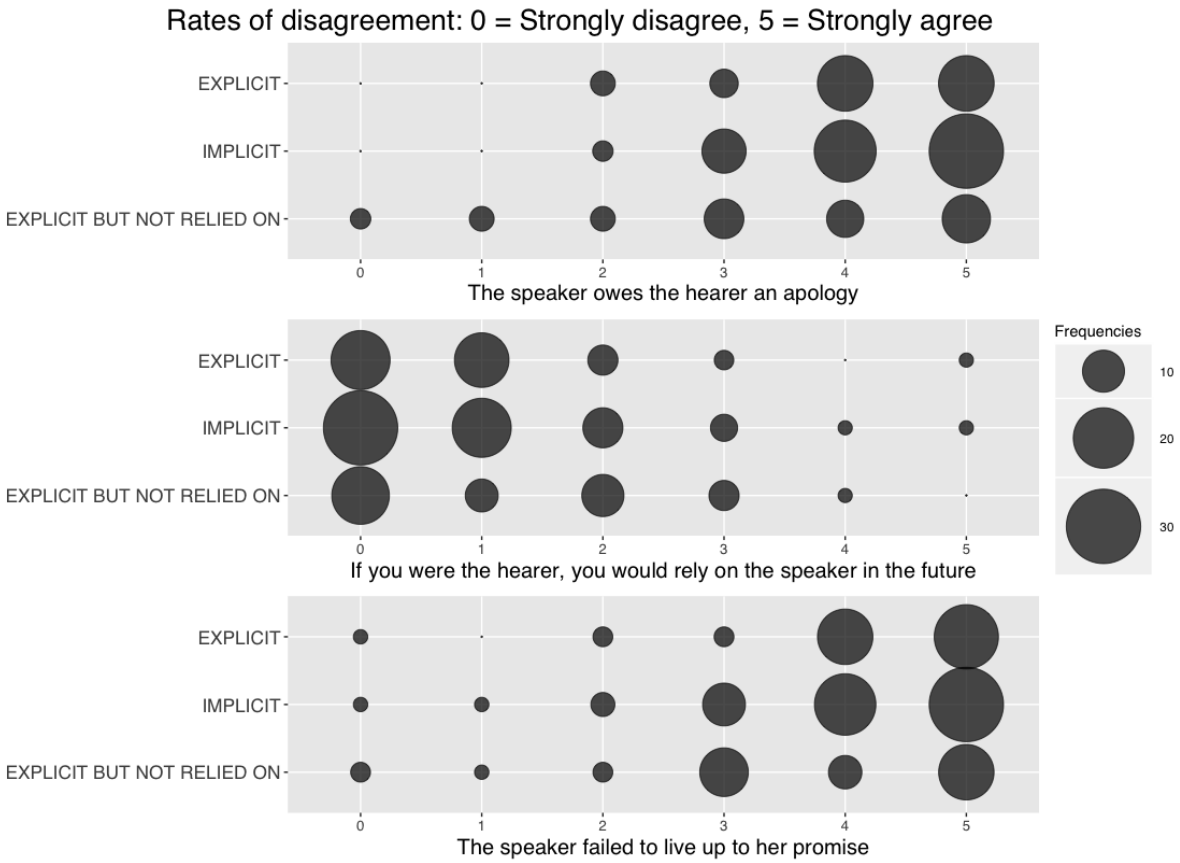

Figure 3.4. Distribution of responses (N = 153) for Study 3c. As can be seen, participants were more likely to judge that an apology was in order in the Implicit condition than in the Explicit but not relied on condition, whereas no difference is found between the Explicit condition and the Implicit condition and between the Explicit condition and the Explicit but not relied on condition. No difference is found between the three conditions for Partner Choice and for Perceived Promise.

3.5 Study 3d

Study 3d replicated the design of the previous studies, but differed in the following respects: (i) the kind of implicit content used is a conversational implicature (a PCIs); and (ii) we presented participants with comic strips rather than with written scenarios or video-clips.

Methods

Participants

Participants were recruited together with participants for Study 3c. From the original dataset, 157 participants were assigned to Study 3d, 54 in the Explicit condition, 57 in the Implicit condition and 46 in the Explicit but not relied on condition. The sample was composed entirely of North Americans. All participants gave their informed consent by ticking a box prior to the experiment.

Materials and procedure

Participants were asked to read different hypothetical situations involving a potential violation of a promise. Subjects were presented with one of the two scenarios described above, in one of three conditions (Explicit, Implicit, Explicit but not relied on) in which the promise is explicit, or only implied, and in which the relevance of the promise is high rather than low.

Subjects were randomly assigned to one of three between-subjects conditions (Explicit, Implicit, Explicit but not relied on). The scenario is presented to be read as if the events occurring in the scenario were true.

In the Implicit condition, the scenario was shown as follows:

> [In a comic strip, X and Y are planning to go home from the office. After crossing a hallway where some free umbrellas are available to be borrowed, X and Y notice that it is raining. While X is about to fetch the umbrellas, Y stops him, stating that: “I am gonna fetch something”. When Y comes back, he only has one umbrella.]

In the Explicit condition, the vignette differed only to the extent that the implied information was now explicitly uttered: "I am gonna fetch two umbrellas". In the Explicit but not relied on condition the vignette was based on the Explicit condition, but differed so that the explicit promise will not be relied upon—specifically, the vignette was changed so that satisfying the statement ("I am gonna fetch two umbrellas") would be irrelevant for the hearer—X and Y would leave the building together, so one umbrella would be enough for them not to get wet (see ***https://osf.io/zrufe/*** for the full vignette).

As in Study 3a and 3b, after reading the vignette, participants were asked to indicate their agreement with some statements. We measured whether the speaker was perceived to have broken a commitment or not. The responses have been collected on a 6-point Likert scale from 0 (strongly disagree with the statement) to 5 (strongly agree with the statement). The commitment measures were the same as in Study 3a and 3b. We again controlled for participants' understanding of the text by asking two comprehension questions (see ***https://osf.io/zrufe/***). Both the comprehension questions and the commitment measures were presented always in a randomised order. Responses from those who failed to answer this question correctly were discarded from the final sample (N = 6).

Results

Similarly to the previous study, to test our hypothesis, we ran a Kruskal-Wallis non-parametric test for each variable.

Consistent with the predictions, a Kruskal-Wallis H test showed that there is a statistically significant difference in the responses to Apology Required [χ^2 (2) = 44.569, p < .001, η^2 = 0.29]. A series of post-hoc pairwise comparison tests showed that the rates of agreement are significantly lower (i.e., speaker is less frequently judged to owe the hearer an apology) in the Explicit but not relied on condition than in the Explicit condition (p < .001), and in the Implicit

condition (p = .006). Contrary to the previous finding, a significant difference is found between the Implicit condition and the Explicit condition (p < .001). Raw data are illustrated in Figure 3.5.

Consistent with the predictions, a Kruskal-Wallis H test showed that there is a statistically significant difference in the responses to Partner Choice, [χ^2 (2) = 52.596, p < .001, η^2 = 0.38]. A series of post-hoc pairwise comparison tests showed that the rates of agreement are significantly higher (i.e., speaker is judged to be significantly more reliable) in the Explicit but not relied on condition than both in the Explicit condition (p < .001), and the Implicit condition (p < .001). However, no significant difference is found between the Implicit condition and the Explicit condition (p = .28). Raw data are illustrated in Figure 3.5.

A Kruskal-Wallis H test showed that there is a statistically significant difference in the responses to Perceived Promise [χ^2 = 46.291, p < .001, η^2 = 0.30]. A series of post-hoc pairwise comparison tests shows that the rates of agreement are significantly lower (i.e. a promise is less frequently judged to be broken) in the Explicit but not relied on condition than in the Explicit condition (p < .001), but significantly higher (i.e. the promise is more frequently judged to be broken) than in the Implicit condition (p = .007). Consequently, a significant difference is found between the Implicit condition and the Explicit condition (p < .001). Raw data are illustrated in Figure 3.5.

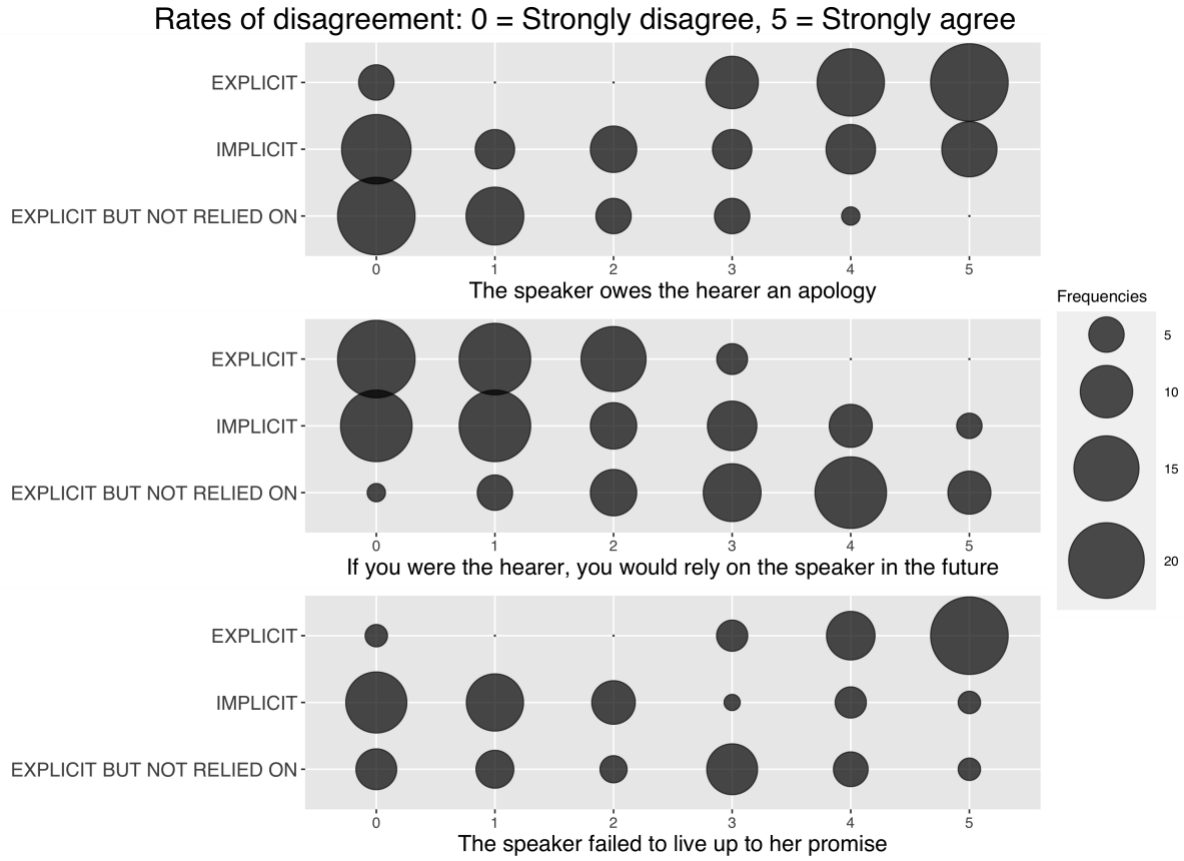

Figure 3.5. Distribution of responses (*N* = 157) for the three measures in Study 3d. As can be seen, participants were more likely to judge that an apology was in order in the Explicit than in the other two conditions, and they were more likely to make this judgement in the Implicit condition than in the Explicit but not relied on condition; they would be more willing to rely on the speaker in the Explicit but not relied on condition than in the other two conditions, whereas no difference is found between the Explicit condition and the Implicit condition; and finally they were more likely to judge that the promise was not lived up to in the Explicit condition than in the other two conditions, and they were more likely to judge that the promise was not lived up to in the Explicit but not relied on than in the Implicit condition.

Discussion

These results corroborate the results of Study 3a and Study 3b (see Table 3.1 and Table 3.2 for summary).

Table 3.1. *Summary of the results of our studies. Our three measures mostly show no differences between how people treated explicit and implicit commitments across the six scenarios.*

The effect of saying-meaning distinction on commitment attribution

	Effect on Apology Required	*Effect on Partner Choice*	*Effect on Perceived Promise*
[Prediction]	*[Explicit = Implicit]*	*[Explicit = Implicit]*	*[Exploratory]*
Study 3a—Rent Scenario	Explicit = Enriched	Explicit = Enriched	Explicit = Enriched
Study 3a—Plant Scenario	Explicit = Enriched	Explicit = Enriched	Explicit = Enriched
Study 3b—Lunch Scenario	Explicit = Implicit	Explicit = Implicit	Explicit > Implicit
Study 3b—Draft Scenario	Explicit = Implicit	Explicit = Implicit	Explicit > Implicit
Study 3c—Video Scenario	Explicit < Implicit	Explicit = Implicit	Explicit = Implicit
Study 3d—Comic Scenario	Explicit > Implicit	Explicit = Implicit	Explicit > Implicit

In Study 3c, as with Study 3b, participants' responses to the commitment questions were not significantly different in the Explicit condition and in the Implicit condition, as predicted (see Table 3.1). On the other hand, contrary to Study 3b, we did not find the predicted difference between the Explicit condition and the Explicit but not relied on condition (see Table 3.2). Rather than reject our hypothesis outright, we reason that the vignette suffers from a lack of salience in the critical difference between these two conditions. Specifically, the presence of additional people at the same table might not have been enough to release the speaker from looking after the laptop, thus (contrary to the intended manipulation) not decreasing the relevance of the promise "I will keep an eye on it." The importance of the tool, of course, increases the type of assurance needed, while the additional people at the table did not offer any. This would mean that *de facto* there was no difference in the relevance of the promise or in the wording between the Explicit condition and the Explicit but not relied on condition.

In Study 3d, as with Study 3a and 3b, we found that breaking a commitment when the hearer relies on the promise has reputational consequences on the violator, and, as with Study 3b and in Study 3c, this is true regardless of whether the promise was made implicitly or explicitly. Speakers were judged to be significantly more reliable in the Explicit but not relied on condition than in the other two conditions, whereas no difference was found between the Implicit condition and the Explicit condition.

We also found that, contrary to Study 3a but consistently with Study 3b, whether the promise was made implicitly or explicitly affected the responses to Apology Required and Perceived Promise questions. On the one hand, speakers were judged to owe the hearer an

apology significantly less frequently in the Explicit but not relied on condition than in the other two, but also less frequently in the Implicit condition than in the Explicit condition; on the other hand, as in Study 3b, the perception of a promise was affected by it being explicitly uttered.

Table 3.2. *Summary of the results of our studies. Our three measures mostly show differences between how people treated Explicit relied on and Explicit but not relied on commitments across the six scenarios.*

The effect of reliance on commitment attribution

[Predictions]	*Effect on Apology Required* *[Explicit > Explicit but not relied on]*	*Effect on Partner Choice* *[Explicit < Explicit but not relied on]*	*Effect on Perceived Promise* *[Exploratory]*
Study 3a—Rent Scenario	Explicit > Explicit but not relied on	Explicit < Explicit but not relied on	Explicit > Explicit but not relied on
Study 3a—Plant Scenario	Explicit > Explicit but not relied on	Explicit < Explicit but not relied on	Explicit > Explicit but not relied on
Study 3b—Lunch Scenario	Explicit > Explicit but not relied on	Explicit < Explicit but not relied on	Explicit > Explicit but not relied on
Study 3b—Draft Scenario	Explicit > Explicit but not relied on	Explicit < Explicit but not relied on	Explicit > Explicit but not relied on
Study 3c—Video Scenario	Explicit = Explicit but not relied on	Explicit = Explicit but not relied on	Explicit = Explicit but not relied on
Study 3d—Comic Scenario	Explicit > Explicit but not relied on	Explicit < Explicit but not relied on	Explicit > Explicit but not relied on

3.6 Discussion of Study 3

In general our experimental results align with our theoretical points (§3.1). Specifically, we found that breaking a commitment when the hearer relies on the promise has reputational consequences for the communicator regardless of whether it is explicitly said or implicitly meant, but not otherwise. Furthermore, reliance on the promise often decisively influenced participants' judgements about whether an apology was necessary or appropriate. Across all our studies people tended to not perceive that a promise that was not relied on (e.g., the implied promise 'I will read it tomorrow' in case the deadline is in two weeks) had even been broken. (The one exceptional finding was that most participants in Study 3b and Study 3d, in cases in which the promises were relied on (e.g., the promise 'I will read it tomorrow' in case the deadline is the day after tomorrow), evaluated the promise primarily taking into account

'what is said' (in line with Saul, 2012). One possibility is that, compared to Study 3a, Study 3b and Study 3d in fact implements a conversational implicature (a GCI rather than a PCI; see §3.4), thus changing perception of what has in fact been said (see Doran et al., 2012).

In sum, our results suggest that audiences take speakers to be committed to 'what is meant' rather than 'what is said' (see below for comments on the generality of this finding). More broadly our findings align with theories of commitment based on social expectations (e.g., Scanlon, 1998; MacCormick & Raz, 1972; Sugden, 2000; for experimental support see also Bonalumi et al., 2019), and they speak against theories which argue that commitments derive from conventions and social practices (e.g., Austin, 1962; Rawls, 1971; Searle, 1969).

These findings raise the interesting question of whether there is a folk concept of promise, tied to 'what is said' rather than to 'what is meant'. There is a small amount of experimental research on the folk concept of lies, which shows that evaluations of whether a lie was uttered depends indeed on what was meant rather than what was said (Weissman & Terkourafi, 2019; Wiegmann et al., 2016; Willemsen & Wiegmann, 2017; Reins & Wiegmann, 2021; see Meibauer, 2018). However, there is not (to our knowledge) any existing experimental research directly focused on the folk concept of promise, and we view this as a productive avenue for future enquiry. Note, incidentally, that if folk intuitions about promises are indeed tied to 'what is said', that creates space for speakers to deny having meant what was only implicitly communicated by invoking this folk beliefs, and hence to avoid commitments that they might otherwise be held to (Lee & Pinker, 2010).

We conclude with some remarks about the potential generality of our findings. Folk intuitions about promises and lies – not to mention other aspects of communication also – are likely to vary between cultures, and maybe within them too. In particular, it is possible that WEIRD societies (Western, Educated, Industrial, Rich, Democratic) will place greater emphasis on what is said, given the history and prevalence of institutions based on what has been made explicit, such as legal systems, formal contracts, and others (Henrich et al., 2010). Future cross-cultural research could explore this hypothesis. At the same time, we expect, looking under the surface of folk intuitions, that commitment to 'what is meant' rather than 'what is said' will *not* exhibit too much variation, either within or between cultures. This is because the underlying cognitive processes involved in communication and commitment attribution are part of the ordinary and robustly developing human cognitive phenotype, which includes an interrelated suite of processes for the expression, recognition, and epistemic evaluation of intentions (Levinson, 2006; Wilson & Sperber, 2012; Sperber et al., 2010; Scott-Phillips, 2015).

Chapter 4. Plausible deniability is affected by partner's reliance

An important challenge in understanding human communication is the question of what processes drive message construction. Why do speakers construct utterances the way they do? How do they generate utterances to achieve their intended effects? Recently, speaker commitment emerged as fundamental in answering these questions (Geurts, 2019): utterances commit the speaker to the truth of a proposition or to a future course of action (Hamblin, 1971; Beyssade & Marandin, 2009), but such commitments are only effective if listeners are able to track these commitments and to hold speakers accountable to them (e.g., Vullioud et al., 2017; see also Mahr & Csibra, 2018, 2020, 2021).

How speakers choose to express their message should impact on how committed they are perceived to be, namely how much others will hold them *accountable* to what they have expressed and the degree of *plausible deniability* of their message. Plausible deniability allows speakers to refute having intended a certain message (typically an implicit one), for instance when confronted by the recipient (Brown & Levinson, 1987; Lee & Pinker, 2010; Pinker, 2007; Pinker et al., 2008). By taking a strategic approach to utterance construction, speakers can manipulate the extent to which the audience can hold them accountable for the meaning they have conveyed (Pinker et al., 2008; Soltys et al., 2014). Illustrating this, Lee and Pinker (2010) found that speakers favoured implicit constructions when they were asked how they would attempt to bribe a policeman. The choice to forego an explicit offer in such a scenario is strategic, as speakers can deny their implicit offer and avoid unpleasant and/or awkward social repercussions. Speakers can thus mitigate the risks of a negative outcome, since 'cooperative' recipients can accept implicit offers, while 'antagonistic' recipients would not have enough evidence to confront them. Therefore, how committed speakers will be perceived to be seems to be critically related to plausible deniability, and how speakers construct their message may in fact be geared towards a modulation of the plausible deniability of their utterance.

However, there are other factors that are likely to modulate perceived speaker commitment beyond how explicitly or implicitly a message is conveyed. For instance, a speaker's confidence in expressing a message impacts both the message's credibility—by increasing its chances of being accepted by the interlocutor—and the speaker's accountability—namely the social repercussions if the message is found to be unreliable (Vullioud et al., 2017; Mazzarella et al., 2018). Similarly, claims about the source of one's information ("I saw it" vs. "Somebody told me") have been shown to have an impact on the message's credibility as well as the speaker's accountability (Mahr & Csibra, 2021).

Extending this work, we want to explore the relationship between two aspects of speaker commitment—accountability and plausible deniability—by focusing on two additional pragmatic factors that may influence speaker commitment: the level of meaning and its strength. First, speakers can convey messages with different degrees of explicitness, that is, using different levels of meaning: speakers communicate not only the linguistically encoded meaning of their utterances, they can also pragmatically *enrich* the content of these utterances or *imply* ('implicate') propositions in addition to what they explicitly say (Grice, 1989). Additionally, speakers can convey messages with different degrees of manifestness—e.g., a pragmatically inferred part of meaning (be it an enrichment or an implicature) may be more or less strongly communicated. Meaning strength is conceived here as the accessibility of what is communicated; it depends both on how manifest the speaker made their intention to communicate a specific content and how important (or inconsequential) the recovery of this content is for the interpretation of the utterance (Wilson & Sperber, 2004). Both the level of meaning and the strength with which a content is expressed may have an impact on how committed a speaker will be perceived to be.

Given that a speaker risks suffering social repercussions when implying something false, would these repercussions be more severe when this content is communicated via a pragmatic enrichment rather than via an implicature? Or would they be worse if the false content is strongly rather than weakly implicated? Alternatively, would the differences in speaker accountability emerge only when these different types of contents – weakly or strongly implicated or enriched – are denied? Here, we aimed to investigate the influence of the level of meaning and meaning strength on the speaker's accountability, and whether they bring about similar effects on their message's plausible deniability.

4.1 How does the level of meaning impact accountability and plausible deniability?

It has recently been proposed that commitment is stronger when meaning is fully linguistically encoded (explicitly communicated) than when it is merely implicated (Morency et al., 2008; Reboul, 2017). Indeed, Mazzarella and colleagues (2018) found that conversational implicatures do not foster as much accountability as explicit contents, while explicit contents and presuppositions lead to similar levels of accountability.

Plausible deniability should be similarly influenced by the degree of explicitness of the conveyed meaning—i.e., by its level of meaning (Brown & Levinson, 1987). Of course, denying

that some fully explicit content was intended is difficult—short of lying or claiming a mistake. The phenomenon of plausible deniability primarily applies to contents conveyed via pragmatic inferences, which seem much easier to revoke. Yet, these do not form a homogeneous category: the outcome of a pragmatic inference might be more or less explicit and thus, we believe, more or less plausibly deniable.

Deniability is closely related to another key notion, that of 'cancellability': all pragmatic phenomena are by definition cancellable—i.e., the utterance that implies a proposition *p* can be followed by the phrase 'but not *p*' without logical contradiction (Grice, 1989; Levinson, 2000). However, not all pragmatic phenomena can be plausibly denied: you may be able to logically cancel part of the content of your utterance without being in a position to plausibly deny intending it in the first place. If, for example, you are asked whether you knew that your young cousin, whom you were babysitting, got into some mischief in the kitchen, and you answer, "I wasn't in the kitchen," the implication that you did not know is clearly cancellable—there is no logical clash between not being present and knowing about something. Yet, it will be hard to deny that you intended to convey your ignorance of the misbehaviour if you are then confronted with the fact that your cousin loudly broke every glass by throwing them at the cat (and missing, thankfully). A deniable proposition is by definition cancellable, but the reverse is not necessarily true: a proposition may be cancellable without being deniable—or at least, not plausibly so (Pinker et al., 2008). Pragmatic phenomena can therefore be denied—which is a start—but whether, and how, they can be plausibly denied is a more complex issue that will depend largely on the properties of the context of utterance (Mazzarella, 2021).

Since Grice (1989), most pragmatists consider there to be different types of implicit (i.e., cancellable) contents. Pragmatic enrichments[15] go beyond what is linguistically encoded in an utterance, but they are linked to linguistic terms or structures (e.g., conjunction *and*, disjunction *or*, quantifiers *some*, *most*), in the presence of which they will often be derived—some theorists even maintain that they are then derived automatically, unless blocked by the context. In

[15] We will refer to pragmatic inferences linked to scalar terms, quantifiers, modals, cardinals, or to logical terms (such as conjunction) as pragmatic 'enrichments' throughout the paper—following authors such as Récanati (2004), Carston (2002) and Sperber & Wilson (1986/1995). These phenomena are considered to be 'Generalised Conversational Implicatures' (GCIs) by Gricean (Grice, 1989) and neo-Gricean theorists (e.g., Horn, 1984; and Levinson, 2000). Although the difference in terminology corresponds to important differences in how different pragmatic theorists view these phenomena, these have no direct bearing on our study.

contrast, implicatures[16] are entirely context-dependent additional propositions and do not depend on any specific linguistic feature. A (rare) theoretical consensus holds that their derivation requires taking into account the proposition explicitly expressed, as well as the context of utterance and the speaker's intention. Because of these differences, enrichments have been at the heart of the debate on the semantic–pragmatic border, with some researchers holding that they are part of the implicit content of the utterance ('what is implicated') (e.g., Horn, 1984), while others—dubbed 'contextualists'—argue that they are part of the explicit content of the utterance ('what is said', in Gricean terms) and contribute to its truth conditions (Récanati, 1993, 2001, 2004; Carston, 2002; Sperber & Wilson, 1986/1995). Theoretical debate notwithstanding, there is a general sense that pragmatic phenomena might be more or less explicit, with implicatures firmly lodged in the implicit camp and pragmatic enrichments verging toward, or achieving, explicitness (see, for example, Levinson, 2000).

A large amount of research focuses on whether participants can distinguish enrichments from explicit content ('what is said' in Gricean terms) to determine whether they contribute to the truth-conditional meaning of the utterance. Enrichments were originally found to be judged as part of 'what is said' (Gibbs & Moise, 1997), but the picture subsequently became more complex (Bezuidenhout & Cutting, 2002). It seems that participants' intuitions differ depending on the type of pragmatic inference (Doran et al., 2009) and the task used—for instance, the types of enrichments investigated by Doran, Ward, Larson, McNabb & Baker (2012) were neither consistently included nor excluded from 'what is said'. Relatedly, several recent studies investigated whether interlocutors consider false information conveyed via pragmatic inferences to be an instance of lying. This research on the comprehension of deceitful implicatures and enrichments yields mixed findings (for a study on the production of misleading enrichment and implicature, see Franke, Dulcinati & Pouscoulous, 2020). Weissman and Terkourafi (2019) show that participants consistently judged false implicatures to be non-lies, while some types of enrichments (e.g., cardinals) were more easily considered to be lies when the inferred meaning was false. Yet, other studies suggest that even false implicatures may be considered eligible to be lies, and sometimes as much so as false enrichments (Antomo et al., 2018; Viebahn et al., 2021; Willemsen & Wiegmann, 2017). Notwithstanding the fact that there are likely cultural differences in people's intuition of what counts as a lie (e.g., Danziger, 2010; Hardin, 2010; Hruschka, 2020), Reins & Wiegmann used a variety of particularly relevant

[16] Similarly, we will refer to the pragmatic inferences corresponding to 'Particularised Conversation Implicatures' (PCIs) in Grice's (1989) terminology simply as 'implicatures'. Note that for most of the relevant experimental literature it would be safe to assume that 'enrichments' and 'implicatures' correspond to the same phenomena as, respectively, GCIs and PCIs.

measures to investigate the folk notion of lying. Following four scenarios involving a false implicature or enrichment, participants were asked whether they considered them to be lies, among other questions; these were compared to an additional set of measures including (among others) explicit questions about commitment (did the speaker commit themselves to the false enrichment/implicature) and deniability (could the speaker convincingly deny it). Lie responses correlated with those assessing commitment and deniability. Reins & Wiegmann found that false implicatures were mostly judged to be lies. However, attributions of lying, as well as commitment and deniability, were lower for implicatures than for enrichments. In a similar vein, Bonalumi Scott-Phillips, Tacha, & Heintz (2020) found that people considered unfulfilled promises conveyed via enrichment to have been broken, but not those conveyed via implicature.

Overall, although enrichments are pragmatic phenomena—and thus cancellable—they often seem to be perceived as contributing to the utterance's truth-conditions. As a result, they generally appear more difficult to deny than implicatures, which are unarguably part of 'what is implicated'. These two levels of meaning should, therefore, modulate both accountability and plausible deniability differently (as suggested by Bonalumi et al., 2020 and Reins & Wiegmann, 2021).

4.2 How does meaning strength impact accountability and plausible deniability?

Another factor that can be linked to both plausible deniability and accountability is the degree of manifestness of what is communicated, i.e., its strength[17]. A pragmatically inferred part of meaning (be it an *enrichment* or an *implicature*) can be more or less strongly communicated; this will depend on how manifest the speaker made their intention to communicate it (Wilson & Sperber, 2002), as well as how essential the pragmatic content is to understand the overall communicative act. First, a strongly pragmatically inferred content (enrichment or implicature) is a proposition the speaker intends to communicate, and they will therefore make this intention clear to their interlocutor, whereas their intention to communicate weaker contents is hazier and, as a result, less manifest to their interlocutor. Second, a strong implicature or enrichment is generally crucial to make the speaker's utterance relevant in context, while the recovery of a weak implicature might be optional. Both aspects

[17] The notion of implicature strength was introduced by relevance theorists (Sperber & Wilson, 1986/1995; 2004; Wilson & Sperber, 2002), yet we believe it to be a very useful tool in the analysis of accountability independently from the specific theoretical apparatus of Relevance theory.

point towards strongly pragmatically inferred parts of meaning as being more accessible for the hearer than weaker ones.

Imagine two young parents arriving at home after braving a downpour and commenting on how they never want to leave the comfort of their living-room again. If one of them exclaims: "We're out of milk!" the implicature that someone must go out to get milk for the child is strongly communicated. Indeed, it is difficult to see how the utterance would be relevant in this context if the implicature was not intended. On the other hand, other (weaker) implicatures might have been intended, or not ("our child drinks more milk than she used to", "we should wean her off milk", "the on-line delivery didn't come on time this week", "you forgot to put milk on the grocery list"…).

As Mazzarella (2021) notes, the strength of an implicit meaning will affect who endorses responsibility for the pragmatic inference. In the case of a strong implicature or enrichment, the speaker bears more responsibility since they make their intention to communicate it clearly manifest. Inversely, in the case of a weak implicature or enrichment the responsibility of deriving it lies mostly with the hearer (Sperber & Wilson, 2006). This, in turn, should involve consequences for how accountable the speaker will be perceived to be: the stronger the implicature or enrichment, the more committed to it the speaker should appear. Plausible deniability should be equally affected by meaning strength. Since the derivation of a strong implicature or enrichment is paramount to understanding the utterance, there is little room left in these cases for an alternative interpretation. Any attempt of denial would, thus, be less plausible, since there is only a narrow range of possibilities to re-construct the context—and thus providing an alternative, non-committal, interpretation of the utterance (Mazzarella, 2021). On the other hand, weak implicatures and enrichments offer the speaker exactly this range of possibilities, suggesting that meaning strength is inversely connected to plausible deniability.

The hypotheses that stronger implicatures and enrichments should be more accessible than weaker ones, but also more commitment inducing, are mostly borne out by the handful of studies investigating meaning strength. Nicolle and Clark (1999) first found that strong implicatures prompted participants to select the implied meaning conveyed by the utterance as the best reflection of 'what [it] said'. In contrast, with weaker implicatures, participants were more likely to select the minimal proposition of the utterance as representative of the explicit content ('what is said'). Consistent with the hypothesis that meaning strength (modulated by relevance) influences accountability and plausible deniability, Bonalumi and colleagues (2020) found that the same explicit broken promise produced different social repercussions for the

speaker depending on whether the recipient was known to rely on the promise made by the speaker. Finally, Sternau, Ariel, Giora, & Fein (2015) investigated the deniability of enrichments, as well as weak and strong implicatures using an explicit question about deniability (akin to Reis & Wiegmann, 2021). Although their findings rely on participants' *a priori* intuitions, rather than actual attempts of denial, they indicated that enrichments are perceived as less deniable than implicatures, and strong implicatures less so than weak ones.

Taken together, these results suggest that strong implicatures are more easily included into the explicit content, have higher impact on accountability and might be harder to deny compared to weaker ones. Note that since meaning strength is a feature of enrichments and implicatures alike (B. Clark, 2013), it should modulate commitment and plausible deniability for both phenomena (Mazzarella, 2021; and Sternau et al., 2017 both also make this prediction for plausible deniability).

4.3 The present study

A message can be conveyed both with different degrees of manifestness, i.e., more or less strongly, and with different degrees of explicitness, i.e., graded levels of meaning (see Figure 4.1). In the present study, to further understand the attribution of commitment (and how speakers may strategically attempt to avoid it), we investigated whether 'meaning strength' and 'level of meaning' modulate accountability and plausible deniability. Specifically, we sought to test the following hypotheses:

- Strongly implicated contents should lead to higher accountability and be more difficult to deny than weakly implicated contents.
- Enrichments should lead to higher accountability, and be more difficult to deny, than implicatures.

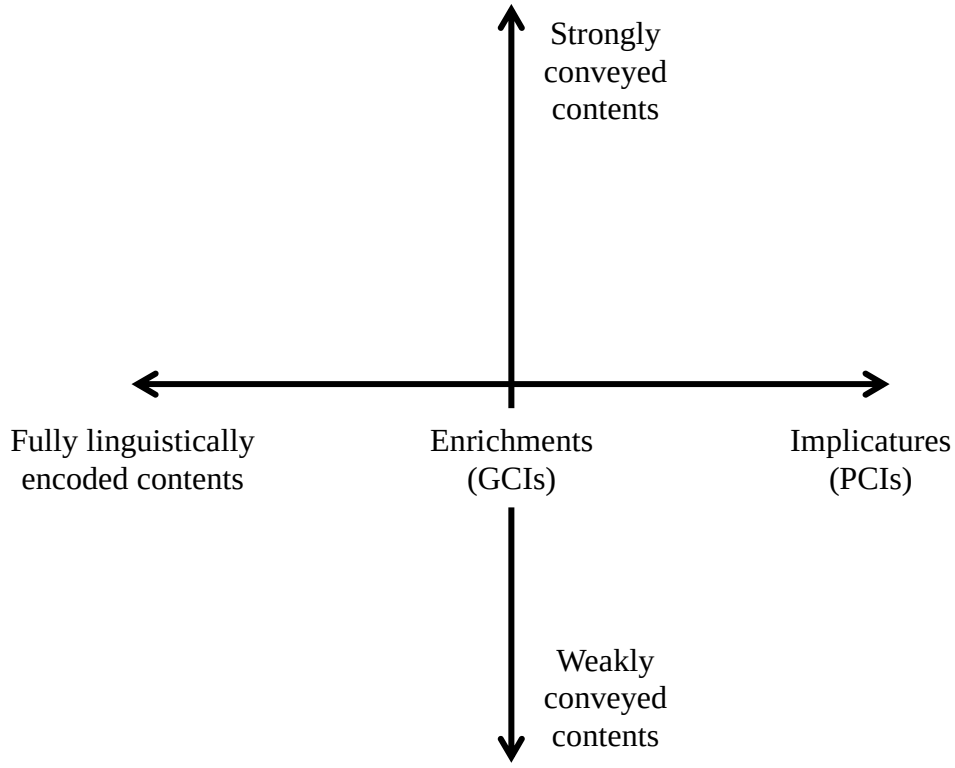

Figure 4.1. The degree of explicitness (x-axis) interacts with the degree of manifestness (y-axis): fully linguistically encoded contents, as well as enrichments and implicatures, can be more or less strongly/weakly communicated.

We tested these hypotheses in an online experiment. The experiment itself—Study 4—was preceded by a norming study to ensure participants interpreted the materials with the pragmatic inferences and drew comparable intended meanings in each experimental condition. Participants were presented with written scenarios in which speakers used different levels of meaning (*Enrichments* or *Implicatures*) with different strengths (*Weak* or *Strong*) to make a commitment. Meaning strength depended on how manifest the implied content was in the scenario. Implicatures were always determined by the interaction between the utterance and the specific context, while enrichments were of four types – the inferences linked to the scalar terms *some* (*not all*) and *or* (meaning *not both*), temporal conjunction enrichment (*and* meaning *and then*), as well as the enrichment of the conditional (*if*) to a biconditional (*if and only if*). We chose four different types of enrichments to make sure our findings could generalise to enrichments as a category and were not limited to a single phenomenon. We took care to pick phenomena generally agreed to be enrichments, or generalised conversational implicatures (GCI), across theoretical frameworks – i.e., all of the phenomena we chose are considered as enrichments or as GCIs by authors such as relevance theorists (Sperber & Wilson, 1986/1995; Carston, 2002; Noveck & Sperber, 2007), Récanati (2004), and by neo-Griceans (such as

Levinson 2000). These authors agree that these pragmatic phenomena differ from (particularised) implicatures and may (unlike implicatures) contribute to the truth-conditions we intuitively assign to the utterances in which they appear.

In the norming study, we measured whether participants inferred the expected enriched/implicated meaning by asking them whether the speaker intended it, as well as their level of confidence about their answer. Study 4 itself was conducted to test our main hypotheses that more strongly implicated contents, on the one hand, and enrichments, on the other, lead to higher accountability and are harder to deny than weaker implicated contents and implicatures, respectively. Participants were presented with scenarios (selected from those tested in the norming study) in which speakers later broke their implied commitment. Following previous studies investigating speaker commitment (Mazzarella et al., 2018; Bonalumi et al., 2020; Yuan & Lyu, 2022), we used two indirect measures to assess whether the speaker was held accountable for breaking their commitment: asking participants whether the speaker was blameworthy and trustworthy. Previous results suggest that blameworthiness and trustworthiness judgments tend to mirror each other (although the former seem to be more influenced by pragmatic factors than the latter, as shown by Mazzarella, et al., 2018). We, therefore, predicted that these two measures would provide similar results: a speaker judged more blameworthy would be deemed less trustworthy, and vice versa. Additionally, the influence of the level of meaning and of meaning strength on denial was also assessed. There is always a risk that participants' *a priori* intuitions on something like deniability (as measured by Reins & Wiegmann, 2021 and Sternau et al., 2015) would be coloured by prejudices. We therefore favoured a more direct approach to establish whether an implied content is perceived as plausibly deniable or not: we compared participants' judgments following the broken commitment with, and without, denial. A decrease of accountability following denial would indicate the plausible deniability of the implied content; if speakers making an implied commitment are held less accountable when it is denied compared to when it is not denied, it is fair to conclude that the implied commitment was (to some extent) deniable. We predicted that the impact of meaning strength and of the level of meaning on plausible deniability would mirror their impact on accountability.

4.4 Norming Study

The norming study focused on the derivation of the appropriate pragmatic meanings, conveyed either by enrichments or by implicatures in different contexts. The results of this study were then used to make informed decisions about which scenarios to use in Study 4.

We expected that most participants would draw the intended pragmatic meaning for both levels of meaning, both in strong and weak contexts. Further, we expected participants to report higher confidence ratings when contents were strongly conveyed than when they were weakly conveyed. Participants were not presented with any breach of the implied commitment, since this might have confounded participants' responses to the implicature question.

Methods

Participants

We recruited 150 participants (105 females, M_{age} = 33.23) through Prolific Academic (Palan & Schitter, 2018). Eligibility criteria for participation in the study were age (20 to 70) and first language (English). We excluded participants if their completion time diverged by more than three standard deviations from the mean (*M* = 379.86 s, *SD* = 375.6 s; *N* = 2), or if they provided wrong answers to the comprehension question (*N* = 29). Participants provided informed consent before taking part in the experiment and were paid £0.84 for their time.

Materials

We created 21 scenarios, each with the following structure:

- Context describing the situation in which the utterance occurs—manipulated so to create *Weak* and *Strong* conditions.
- Dialogue containing the speaker's implied commitment to accomplish a specific task. The commitment was conveyed either via *Enrichment* or *Implicature* (see Table 4.1). All the scenarios are available at ***https://osf.io/2fu93/***.

Table 4.1. *Examples of scenarios in the four conditions.*

Enrichment condition	
Weak condition	*Strong condition*
Sophie and Elliot are colleagues and both work in a bar as bartenders. There is no more craft beer on tap and Sophie and Elliot have to change the keg. It is a Tuesday night and there are very few customers.	Sophie and Elliot are colleagues and both work in a bar as bartenders. There is no more craft beer on tap and Sophie and Elliot have to change the keg. It is a Saturday night and there are a lot of impatient customers.
Sophie: There's no more craft beer on tap. Elliot: **I'll finish making this cocktail and change the keg.**	
Implicature condition	
Weak condition	*Strong condition*
Arthur is looking for someone to babysit his daughter tomorrow night, because he has a work event to attend. He is talking about his predicament with his friend, Olivia. They both know that Olivia's work has very unpredictable hours. She has cancelled plans with Arthur at the last minute in the past.	Arthur is looking for someone to babysit his daughter tomorrow night, because he has a work event to attend. He is talking about his predicament with his friend, Olivia. Olivia has babysat Arthur's daughter in the past and he found her reliable.
Arthur: I don't know if I'll be able to find someone before tomorrow. I don't know what to do. Olivia: **I'm free tomorrow night.**	

We modulated meaning strength through changes in the scenario context resulting in an increase or decrease of the relevance of the speaker's utterance. This change in relevance should affect the accessibility of the implied commitment. For instance, in the scenario mentioned in Table 4.1, the change in context would concern the number of customers waiting to be served. As Elliot utters, "I'll finish this cocktail and change the keg," in a weak context where few customers are waiting, the order in which he does so should be less relevant than in the context in which there are numerous customers to serve. Although the implicature would be drawn in both contexts, in the latter, stronger context, Sophie would understand Elliot's utterance as more pointedly ordered, thus resulting in a more accessible and stronger implicature.

The level of meaning was examined using 8 scenarios conveying commitment via an implicature and 13 scenarios conveying commitment via an enrichment. Four types of enrichments were used: two types of scalar implicatures—one linked to the quantifier 'some' (*some*, but *not all*), the other to the connective 'or' (but *not both*) —as well as the temporal enrichment of the connective 'and' (to *and then*) and conditional perfection, where *only if* is

derived from 'if' (see Table 4.2; for discussion of these phenomena see for instance, Levinson, 2000; Carston, 2002; Noveck, 2004).

Table 4.2. *Examples of different types of enrichment used in the norming study and in Study 4.*

Type of enrichment	Example	Intended meaning
Scalar implicature quantifier 'some'	I'll take **some** tissue packets with me to the office today.	I'll take **some** tissue packets with me to the office today, **but not all** the tissue packets.
Scalar implicature connective 'or'	Yes, I'll need the projector **or** the eraser board.	Yes, I'll need **either** the projector **or** the eraser board, **but not both**.
Conjunction enrichment of 'and'	I'll finish making this cocktail **and** change the keg.	I'll **first** finish making this cocktail **and then** change the keg.
Conditional perfection with 'if'	**If** you don't get a dog **then** I'll get a cat	I'll get a cat, **if and only if** you don't get a dog.

Procedure and Design

The norming study used a mixed factorial 2 ('strength' as between-subjects factor: *Strong* vs. *Weak*) x 2 ('level of meaning' as within-subject factor: *Enrichment* vs. *Implicature*) design. Participants were randomly presented with four enrichments and two implicatures and saw all of them either in a strong or in a weak version. After reading each scenario, participants answered a comprehension question about the scenario content. Participants were then reminded of the speaker's utterance containing the implicit commitment, and were asked to answer an Implicature question and a Confidence Rating question, as illustrated below:

- Implicature Question:

Do you understand Elliot [*speaker*] to have meant that he will finish making the drink first, and then change the keg [*implied meaning*]?

(Yes/No/I don't know)

- Confidence Rating:

How confident are you in your answer? Rate your confidence from 0 to 5, with 0 being 'not at all' and 5 being 'completely'. (6-point Likert scale)

Results and Discussion

We took the answers 'No', and 'I don't know' to the implicature question as indicating that the implied commitment had not been inferred by participants, while 'Yes' answers were interpreted to indicate that the implicature had been drawn. To get a measure of how reliably participants inferred the implied meaning, we converted participants' responses to the implicature and confidence questions into a unified 'inference score' by multiplying participants' confidence responses by -1 when participants failed to draw the correct inference and by +1 when they succeeded (for a similar approach, see Starmans & Friedman, 2012).

Based on this measure, we selected eight target scenarios to be used in Study 4: four scenarios in which the commitment was conveyed through an implicature and four scenarios in which the commitment was conveyed through an enrichment, one in each category: quantifier 'some', connectives 'or' and 'and', and conditional 'if'. Selected scenarios strongly differed from excluded ones in terms of their inference scores ($M_{Selected}$ = 4.18, $SD_{Selected}$ = 0.59; $M_{Excluded}$ = 2.78, $SD_{Excluded}$ = 1.23) as determined by a two sample t-test (t(38.19) = 4.95, p < .0001).

Next, we analysed selected scenarios to test whether 'meaning strength' and 'level of meaning' had an impact on participants' ability to draw the correct pragmatic inference. To do so, we fit a linear mixed-effects model with scenario as random intercept (with fixed slopes) to each participant's inference score. As fixed effects, we included 'meaning strength' (*Weak/Strong*) and 'level of meaning' (*Enrichment/Implicature*). We did not include varying slopes for our random effect because a model including such slopes resulted in a singular fit. We estimated the effect of each of these factors via Satterthwaite approximations of degrees of freedom (Luke, 2017). For parameter estimation, we generated 95% confidence intervals around beta-values using parametric bootstrapping. These analyses (and the ones below for Study 4) were performed using *R* version 3.4.1 (R Core Team, 2020), and *RStudio* (RStudio Team, 2020) with the packages *lme4* (Bates et al., 2015) and *lmerTest* (Kuznetsova et al., 2017). For a discussion of the trend towards mixed-effects modelling in the behavioural and neural sciences see Boisgontier and Cheval (2016).

This analysis showed that meaning strength (β = -0.59, SE = 0.23, t = -2.54, p = .039, 95% CI = [-1.02, -0.19]) but not the level of meaning (β = -0.21, SE = 0.3, t = -0.70, p = .508, 95% CI = [-0.79, 0.38]) impacted participants' inference scores.

Consistent with our prediction, participants' inference scores were lower in *Weak* scenarios (M = 3.88, SD = 0.68) than in *Strong* scenarios (M = 4.47, SD = 0.27; see Figure 2).

Critically, we did not find differences in the proportion of participants who inferred the intended meaning in both types of scenarios, $t(9.7) = -1.97$, $p = .078$; 92.96% of participants inferred the intended meaning in *Weak* conditions and 98.34% inferred the intended meaning in *Strong* conditions. Thus, while meaning strength did not affect the likelihood of the inferences themselves being drawn, it impacted how confident the participants were in these inferences, $t(7) = 3.05$, $p = .018$.

Further, participants achieved similar inference scores in scenarios with *Enrichments* ($M = 4.29$, $SD = 0.44$) and scenarios with *Implicatures* ($M = 4.08$, $SD = 0.37$; see Fig. 4.2), suggesting that the level of meaning affected neither the likelihood of the inferences themselves being drawn, nor how confident the participants were in these inferences.

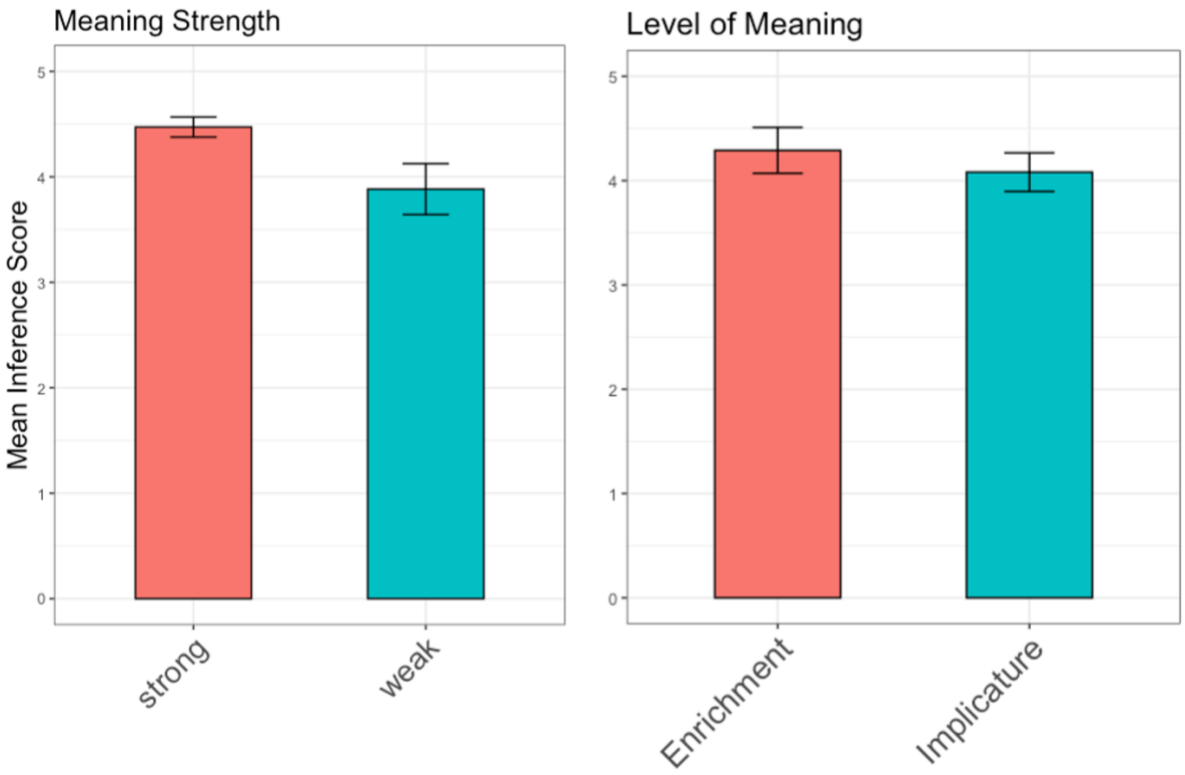

Figure 4.2. Mean inference scores for different meaning strengths (left) and levels of meaning (right) in the norming study. Error bars depict standard errors. Whereas inference scores differed between Strong and Weak conditions, they did not differ between levels of meaning.

The norming experiment showed that participants robustly derived the intended meanings in our selected scenarios. Moreover, while the level of meaning did not affect the derivation nor the confidence of these inferences, the strength of the implied meaning did affect the

confidence with which participants drew these pragmatic inferences as predicted. In light of these results, we therefore expected meaning strength to play a particularly prominent role in modulating accountability and enabling plausible deniability in Study 4.

4.5 Study 4

Study 4 was designed to test our main hypothesis that meaning strength and level of meaning modulate both the extent to which a speaker is held accountable for their implied meaning and the extent to which a speaker can plausibly deny such meaning. We predicted that contents conveyed via enrichment would lead to higher accountability ratings and would be less easily deniable than contents conveyed via implicature. We also predicted that strongly conveyed contents would lead to higher accountability ratings and would be less easily deniable than weakly conveyed contents.

We measured perceived accountability through two test questions: one asking how blameworthy the speaker is (blame question), the other asking how much participants would trust the speaker in the future (mistrust question). The blame question measured commitment attribution via blameworthiness; the mistrust question measured whether participants would draw some form of trait-inference relevant to future interactions resulting from the breach of the commitment. While accountability was measured using direct questions, to assess the effect of plausible deniability we introduced an additional between-subjects factor: the denial of the implied commitment. If the accountability ratings of an implied commitment are lower when it is denied than when it is not, then the implied commitment was (at least to some extent) deniable.

Methods

The study was pre-registered on OSF.io, with the sample size, planned analyses and participant exclusion criteria specified. The pre-registration document is available at ***https://osf.io/nkv63/***.

Participants

A power analysis conducted with G*Power 3.1 (Faul et al., 2009) revealed that, assuming a small to medium effect size (F = 0.175), a sample size of 400 participants would convey a statistical power of 0.89. Consequently, we recruited 400 participants (297 females, M_{age} = 35.39) through Prolific, with age (between 20 and 70) and first language (English) as eligibility criteria. Participants provided informed consent and were compensated with £0.09. Data was

discarded for participants who took too much or too little time to complete the study (three standard deviations from the mean; *N* = 7) or failed the comprehension question (*N* = 6), totalling 13 excluded participants. The final analysis thus included 387 participants.

Materials

Study 4 used the eight scenarios selected based on the results of the norming experiment. In four scenarios the meaning was conveyed through an implicature and in four others the meaning was conveyed through an enrichment, one in each category: quantifier 'some', connectives 'or' and 'and', conditional 'if' (see, Table 4.2).

In contrast to the norming experiment, Study 4 included an additional between-subjects factor: 'denial'. In the *Denial Present* condition, the breach of the commitment was followed by the speaker's denial of the implied content, when confronted by the recipient. The denial always appeared in the following form: "I didn't say I would *[implied commitment]*. I said I would *[explicit utterance]*" (see Table 4.3 for an example).

The scenarios had the following structure:

- Context describing the situation in which the utterance occurs—manipulated so to create *Weak* and *Strong* conditions.
- Dialogue containing the speaker's implied commitment to accomplish a specific task—conveyed either via *Enrichment* or via *Implicature*.
- Breach of the implied commitment.
- Dialogue containing the speaker's denial of the implied commitment—only the participants in the *Denial Present* condition were presented with this dialogue.

In contrast to the norming experiment, there was no implicature question in the main experiment, since the presence of denial in half of the conditions would have confounded the responses to the implicature question.

Design and procedure

Study 4 had three between-subjects factors: 'strength' (*Strong* vs. *Weak*), 'level of meaning' (*Implicature* vs. *Enrichment*), and 'denial' (*Denial Present* vs. *Denial Absent*), resulting in eight different experimental conditions. Participants were randomly assigned to one of the eight conditions, and each participant was only exposed to one scenario in one condition. Participants were told to read the scenario, and were then presented with a multiple-choice comprehension question. Participants were subsequently reminded of the speaker's utterance

containing the implied commitment and were presented with a blame question and a mistrust question on 6-point Likert scales (see Table 4.3 for an example).

Table 4.3. Scenario structure and measures used in Study 4 with example (of the Strong + Enrichment + Denial Present conditions).

	Example
Context *(either Strong or Weak)*	Sophie and Elliot are colleagues and both work in a bar as bartenders. There is no more craft beer on tap and Sophie and Elliot have to change the keg. It is a Saturday night and there are a lot of impatient customers.
Implied commitment *(either Enrichment or Implicature)*	Sophie: There's no more craft beer on tap. Elliot: I'll finish making this cocktail and change the keg.
Breach of implied commitment	Elliot leaves the cocktail he was making and goes to change the keg. The customer whose cocktail it was complains to Sophie. Sophie is unhappy about this.
Comprehension Question	Where are Sophie and Elliot working? • A café • A pub • A restaurant • An ice-cream parlour
Denial of implied commitment *(either Present or Absent)*	Sophie: You said you would finish making the cocktail before changing the keg! Elliot: I didn't say that I would do that first. I said I'd finish making the cocktail and change the keg.
Utterance reminder	Remember Elliot said: Elliot: I'll finish making this drink and change the keg.
Blame Question	If you were *Sophie [listener]* how much would you blame *Elliot [speaker]* for misleading you? Rate your confidence from 0 to 5, with 0 being 'not at all' and 5 being 'completely'. [Likert scale]
Mistrust Question	If you were *Sophie [listener]* how much would you mistrust *Elliot [speaker]* in the future? Rate your confidence from 0 to 5, with 0 being 'not at all' and 5 being 'completely'. [Likert scale]

We expected the blame question and the mistrust question to show similar patterns of results. For the blame question, we predicted that participants would be more likely to blame

the speaker in the *Strong* condition than in the *Weak* condition, and that they would be more likely to blame the speaker in the *Enrichment* condition than in the *Implicature* condition. Similarly, for the mistrust question, we predicted that participants would be more likely to mistrust the speaker in the *Strong* condition than in the *Weak* condition, and that they would be more likely to mistrust the speaker in the *Enrichment* condition than in the *Implicature* condition.

Moreover, we predicted interactions between meaning strength and the presence of denial, as well as between level of meaning and the presence of denial. First, we expected weakly (but not strongly) conveyed contents to be more plausibly deniable. Therefore, we predict that in the *Weak* condition (but not in the *Strong* condition) participants would be less likely to blame and to mistrust the speaker in the *Denial Present* condition compared to the *Denial Absent* condition. Second, we expected implicatures (but less so enrichments) to be more plausibly deniable. Therefore, we predicted that in the *Implicature* conditions (but not in the *Enrichment* condition), participants would be less likely to blame and mistrust the speaker in the *Denial Present* condition compared to the *Denial Absent* condition.

Results

Figures 4.3-4.5 illustrate the distribution of responses across experimental conditions. To test our hypotheses, we computed two separate linear mixed-effects models for each of our dependent variables (blame and mistrust ratings) using 'scenario' as a random intercept (with fixed slopes). Just as in the norming experiment, significance values were computed by computing degrees of freedom with Satterthwaite approximations. Factors for fixed-effects were entered into the model via sum-coding. For parameter estimates, we generated 95% confidence intervals via parametric bootstrapping. The results of this approach are summarized in Table 4.4.

Across participants and conditions, blame and mistrust responses were significantly correlated ($r = 0.62$, $p < .001$). Nonetheless, these two measures were not equally sensitive to our experimental manipulations.

Table 4.4. *Results of the linear mixed-effects analysis for the blame and mistrust measures of Study 4. Effects significant at the 0.5 level are printed in bold. Non-significant interaction effects were dropped from the final model and are therefore not displayed here.*

Outcome	Model	β	*SE*	*t*	*95% CI*	*p*
Blame	Random Intercept (Scenario)	0.01	0.13	0.04	-0.25 – 0.26	.97
	Denial	-0.05	0.05	-1.09	-0.14 – 0.03	.279
	Strength* **	**0.20**	**0.05**	**4.29**	**0.11 – 0.29**	**<.0001**
	Level of Meaning	-0.07	0.13	-0.53	-0.29 – 0.17	.615
	Denial x Strength*	**0.12**	**0.05**	**2.56**	**0.03 – 0.21**	**.011**
	Denial x Level of Meaning**	**0.14**	**0.05**	**3.07**	**0.06 – 0.24**	**.002**
Mistrust	Random Intercept (Scenario)	0.001	0.13	0.01	-0.26 – 0.27	0.99
	Denial	0.06	0.05	1.15	-0.04 – 0.15	.252
	Strength	0.02	0.05	0.51	-0.07 – 0.12	.612
	Level of Meaning	-0.06	0.05	-0.46	-0.34 – 0.23	.66
	Denial x Level of Meaning*	**0.12**	**0.05**	**2.39**	**0.01 – 0.2**	**.018**

For the blame measure we found, as predicted, a significant effect of strength, indicating higher blame ratings in *Strong* (*M* = 3.8, *SD* = 1.14) than in *Weak* conditions (*M* = 3.28, *SD* = 1.36; see Figure 4.3). We did not find a main effect of denial, nor of level of meaning on blame responses.

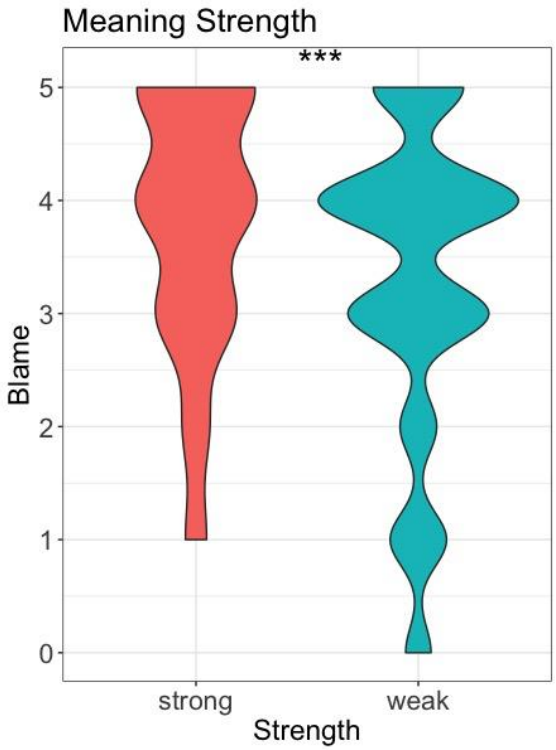

Figure 4.3. Violin plots of distributions of blame ratings in Weak and Strong conditions. Blame ratings were lower in Weak than in Strong conditions.

However, as predicted, a significant interaction between denial and level of meaning showed that *Implicature* conditions were associated with lower blame rates in the *Denial Present* condition (*M* = 3.86, *SD* = 1.27) compared to the *Denial Absent* condition (*M* = 3.38, *SD* = 1.29). In contrast, *Enrichment* conditions were not affected by the presence of denial. Finally, we found an interaction between denial and strength: as predicted, *Weak* conditions were associated with lower blame rates in the *Denial Present* condition (*M* = 3.07; *SD* = 1.23) than in the *Denial Absent* condition (*M* = 3.51; *SD* = 1.44). In contrast, the presence of denial did not influence blame rates in *Strong* conditions. Figure 4.4 displays these interaction effects on blame rates.

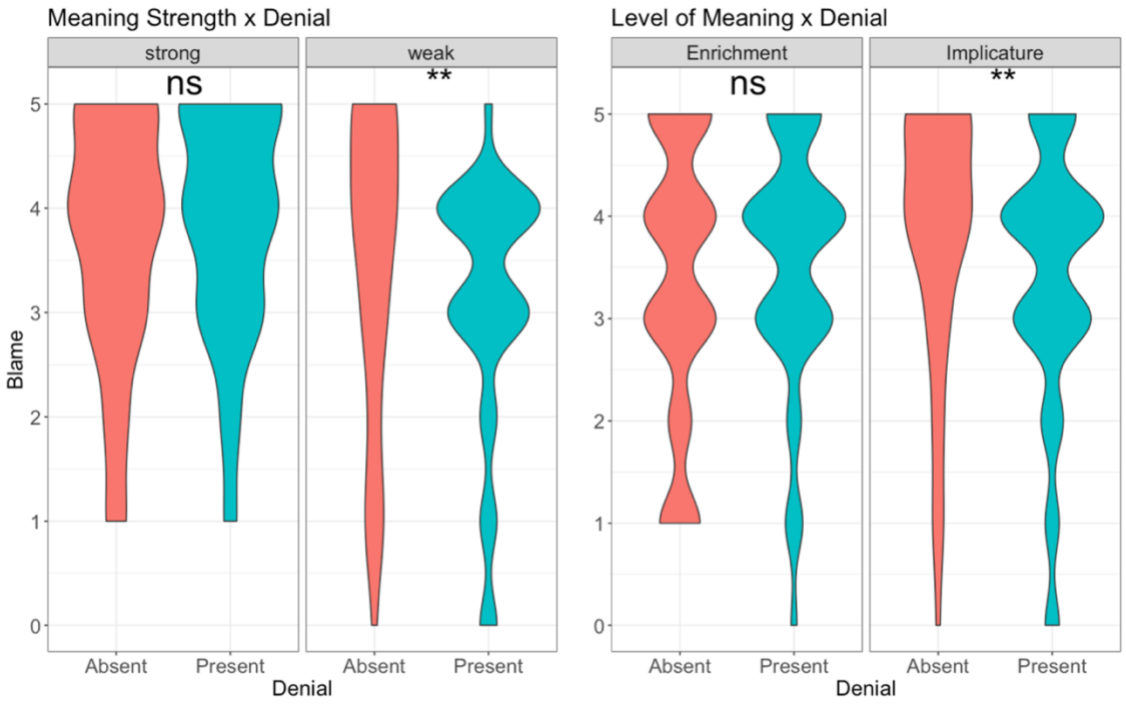

Figure 4.4. Violin plots of distributions of blame ratings across meaning strengths, presence of denial, and levels of meaning. Denial was associated with lower blame ratings in Weak but not Strong conditions. Further, denial was associated with lower blame ratings when meaning was conveyed via Implicature but not when meaning was conveyed via Enrichment.

For the mistrust measure, however, we found only a significant interaction between denial and level of meaning (see Table 4.4 and Figure 4.5), indicating that enrichments, but not implicatures, were associated with higher mistrust in *Denial Present* conditions (M = 3.09, SD = 1.41) than the in the *Denial Absent* conditions (M = 2.6, SD = 1.5). All other comparisons were not significant for the mistrust measures.

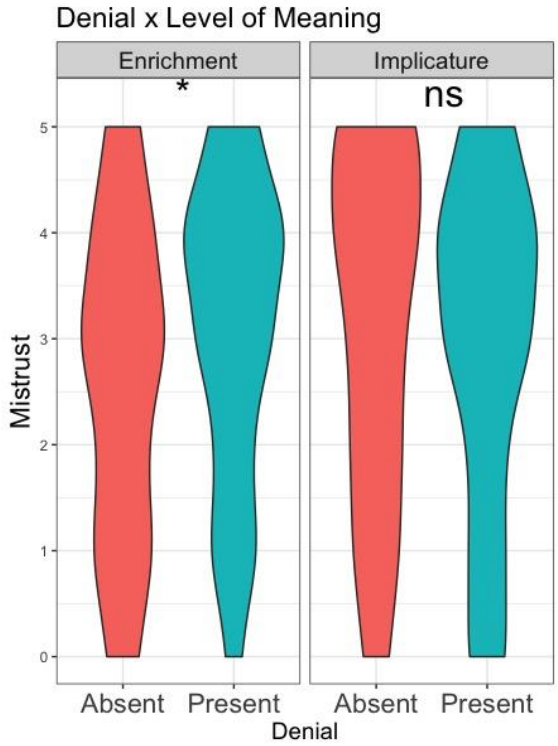

Figure 4.5. Violin plots of mistrust ratings in Denial Present vs. Denial Absent and Enrichment vs Implicature conditions. Denial was associated with higher mistrust ratings in Enrichment but not in Implicature conditions.

These findings indicate that both meaning strength and level of meaning influence plausible deniability: the presence of a denial led to a decrease of speaker's blameworthiness only when the implied content was weakly rather than strongly conveyed, and/or this content was conveyed via implicature rather than enrichment. The presence of denial also led to an increase of speaker's perceived lack of trustworthiness when this content was an enrichment. Furthermore, across all conditions the breach of strongly conveyed commitments led to higher perceived blameworthiness than weakly conveyed commitments. Even though the mistrust measure was less sensitive than expected, our predictions were largely borne out.

4.6 Discussion of Study 4

During the 2016 American presidential election campaign, Republican candidate Donald Trump often mentioned his plan to build a wall along the US-Mexico border, along with the idea that Mexico would take on the economic responsibilities for this operation. For instance, in his presidential announcement speech, he declared "[…] I will build a great, great wall on our

southern border. And I will have Mexico pay for that wall"[18]. In January 2020, President Trump commented to reporters at the White House[19]: "I didn't say they're going to write me a check […]. I said they're going to pay for the wall. And if Congress approves this incredible trade bill[20] that we made with Mexico […] they are paying for the wall many, many times over. […] When I say Mexico is going to pay for the wall, *that's* what I mean [emphasis added], Mexico's paying for the wall". In this, as in other circumstances, Trump (and many before him) exploited an intuitive aspect of communication: whenever a commitment is broken (a promise is not lived up to, or an assertion is found to be false), a good strategy to limit social repercussions is to deny having committed to it in the first place. A good strategy to refute a commitment is to deny the alleged meaning of the original utterance ('I didn't say that…') or to appeal to an alternative meaning ('I meant that…') (Boogaart et al., 2020). Pinker and colleagues have argued that speakers in certain circumstances capitalise on these options when they generate their messages (Pinker et al., 2008; Lee and Pinker, 2010).

The current study investigated whether implicitly conveyed commitments are indeed plausibly deniable, i.e., whether speakers can get out of such commitments by denying they intended the alleged meaning of their utterance. If a commitment is plausibly deniable the costs the speaker incurs following its breach will be mitigated after denial. However, the implicit/explicit dichotomy is not the only factor at play. Our design therefore explored the impact of two factors: the degree of explicitness (the level of meaning) of what is communicated and its strength.

Across all experimental conditions, the breach of strongly conveyed commitments led to higher blameworthiness than weakly conveyed commitments. Furthermore, our findings indicate that both meaning strength and level of meaning influence plausible deniability: in the presence of a denial, participants judged the speaker to be less blameworthy when the implied content was weakly conveyed (but not when it was strongly conveyed), or when this content was an implicature (but not an enrichment). This pattern was also partly reflected in judgments about speaker's untrustworthiness: the presence of a denial led to an increase in untrustworthiness when the implied content was an enrichment (but not an implicature). In

[18] ***https://www.washingtonpost.com/politics/2019/live-updates/trump-white-house/live-fact-checking-and-analysis-of-president-trumps-immigration-speech/a-history-of-trumps-promises-that-mexico-would-pay-for-the-wall-which-it-refuses-to-do/***. (The Washington Post, 2019 Jan 9).

[19] ***https://edition.cnn.com/2019/01/10/politics/trump-mexico-pay-wall/index.html*** (CNN, 2019 Jan 10).

[20] The United States-Mexico-Canada Agreement (USMCA), approved by US Senate and signed into law on January 2020 (H.R. 5430 — 116th Congress: United States-Mexico-Canada Agreement Implementation Act., 2022).

view of the results of the norming study it is unlikely that these observations could be attributed to participants failing to draw the intended implicatures from our materials.

As predicted, our results show that denial works better in weak contexts than in strong contexts: weak contents showed higher deniability (i.e., they were associated with lower blame ratings after denial) than strong contents, for which blame rates remained unaffected by denial. This is one of the key findings of our study. The fact that weakly conveyed contents are easier to deny is consistent with Nicolle and Clark (1999), who found that strongly conveyed meanings (in contrast to weakly conveyed ones) are frequently associated with 'what is said'. Our results also reinforce the conclusions of Sternau et al. (2015): when explicitly asked, their participants found implicatures more deniable than enrichments, and weak implicatures more deniable than strong implicatures. Together, these findings confirm that, as argued by Mazzarella (2021), meaning strength exerts a considerable influence on plausible deniability.

Importantly, the current study also shows that levels of meaning modulate plausible deniability. Implicatures appeared to be more easily deniable (with blame rates lower after denial) than enrichments (blame rates were unaffected by denial). This is reminiscent of Reins & Wiegmann (2021), who found a similar pattern relying on participants' *a priori* intuitions about denial. We used four different types of enrichments—scalar inferences linked to the quantifier *some*, and to disjunction, conjunction enrichment, as well as conditional perfection; our results are therefore not reduceable to a single pragmatic phenomenon, instead they can be, to some extent, generalised to enrichments in general.

These findings are also consistent with the contextualist claim that enrichments can be included into the truth-conditions of the utterance (i.e., 'what is said') (Récanati, 1993, 2001, 2004; Carston, 2002; Sperber & Wilson, 1986/1995). If enrichments affect the truth-conditional content of the utterance and are perceived as being part of 'what is said', then they should, *ceteris paribus*, be harder to deny. Our results are thus in line with the those of Doran et al. (2012) indicating that, contrary to implicatures, enrichments are sometimes included in 'what is said'; Weissman and Terkourafi (2019), who show that false enrichments are considered lies, while false implicatures are not (see below for further discussion); as well as Bonalumi et al. (2020), who highlight that unfulfilled enriched promises are considered *de facto* broken promises, while unfulfilled implicated promises are not.

Importantly, and as predicted, our results also show that meaning strength impacts accountability: the breach of strongly implied commitments was associated with higher blame rates compared to the breach of weakly implied commitments –both in the presence and the absence of denial. Since meaning strength was modulated through changes of the context that

affected the relevance of what was communicated, our results suggest that the relevance of the conveyed content modulates the social consequences of breaches of commitments.

Contrary to our hypotheses, however, we did not find any influence of level of meaning alone on participants' blame rates. This is surprising, considering the theoretical and experimental background, which would suggest an overall difference with regards to accountability between different levels of meaning. We collected two measures of accountability: a blameworthiness judgment and an untrustworthiness judgment. While the two judgments were correlated, mistrust judgments were differently affected by our experimental manipulations. Meaning strength did not impact participants' mistrust judgments in either the presence or in the absence of the denial. One possible reason for this discrepancy might be that the mistrust question was always presented after the blame question. It may be that some participants, after having simulated a partner-control strategy (i.e., blaming the speaker for falsely implying something), perceived an additional partner-choice strategy (i.e., mistrusting the speaker) as redundant or excessively severe (see Noë & Hammerstein, 1994). It may also be the case that blameworthiness is more severely impacted by meaning strength than trustworthiness. It could also be that the multiple utterances presented in the stimuli were themselves carrying disparate illocutionary force, causing an array of other expectations that were instead (partially) satisfied. There would then be no reason left to consider speaker untrustworthy, despite still being blameworthy for the misleading utterance. One last possibility concerns the phrasing of the mistrust question, which may have been confusing for the participants (namely, they could have interpreted it as a *trust* question). On the other hand, our untrustworthiness measure reveals one significant interaction: when the speaker attempted to deny a commitment conveyed by an enrichment, the speaker's trustworthiness was jeopardised compared to situations when the speaker did not attempt to deny it (in line with Mazzarella et al., 2018). It may be that certain types of level of meanings such as enrichments are considered lies, and as such they are not only harder to deny, but possibly not deniable at all, calling for a more radical strategy such as mistrusting the speaker in the future.

Interestingly, some very recent studies have linked the concept of lying to commitment (Reins & Wiegmann, 2021; Wiegmann et al., 2021). Specifically, they argue that participants' intuitive understanding of lying does not only encompass intentionally false assertions (part of 'what is said'), but also false implicatures to which the speaker seems committed (against

traditional views, which restrict lying to what is explicitly asserted, see, e.g., Carson, 2006, 2010; Saul, 2012; Stokke, 2018).[21]

Claims concerning the nature of lying and whether it should include misleading implicit communication are beyond the scope of our work. Nevertheless, our findings have interesting consequences for this wave of thinking, which equates lying with misleading content that is communicated intentionally, including implicit content—as long as the speaker appears sufficiently committed to it. With commitment becoming paramount also for our assessment of lying, it is essential to elucidate the markers of commitment. In a sense, the present study precisely highlights two factors routinely used to modulate commitment: the speaker's choice of level of meaning and the strength of the implicit content. Indeed, the difference we find between enrichments and implicatures, both in terms of accountability and of deniability, echoes the results of Reins & Wiegmann (2021). Even more crucially, our findings emphasize the role of meaning strength in how committed the speaker appears to be to a given implicit content: meaning strength seems to be the factor that shape our intuitions about commitment most effectively. If speaker commitment is the key to the hearer's perception of truth and lying – and their related social consequences – it is critical to further investigate the role of meaning strength.

When speakers aim to limit their commitment and the social consequences of communicating potentially unwelcome or unreliable messages, one good strategy is to increase the room for plausible deniability (Lee & Pinker, 2010; Pinker, 2007; Pinker et al., 2008). The results presented here provide clear evidence of the effective impact of a denial on speaker's accountability. Our findings show that accountability can only be mitigated by denial depending on certain pragmatic factors, specifically the level of meaning and the strength of what is communicated: despite being logically cancellable, not all contents conveyed implicitly by the speaker can be plausibly denied. To strategically manipulate their accountability, speakers must therefore not only use implicit formulations, but also carefully balance the level of implicitness and, crucially, the strength, of what they communicate implicitly. Only when these factors are properly intertwined can speakers optimally minimise the social consequences of potentially unreliable messages. Strategic communication is a complex phenomenon that goes far beyond the implicit/explicit dichotomy.

[21] As we mention in the Introduction, in some studies participants do not consider most deceptive implicatures as lies (Weissman & Terkourafi, 2019), while they appear to in others (Antomo et al., 2018; Or et al., 2017; Reins & Wiegmann, 2021; Wiegmann et al., 2021; Willemsen & Wiegmann, 2017).

Part III. Children's reactions to commitment violations

The previous chapters focused on how adults react to commitment violations in joint and communicative contexts. Such violations were recognised and caused normative reaction despite the fact that commitments were only implicitly cued. Are young children similarly sensitive to commitment violations in joint and communicative contexts? When are children starting to develop a sensitivity to the normative consequences of commitments? In this next part I will present two studies investigating children's understanding of the scope of the commitment. Do children recognise when these normative consequences hold and when they do not? Do children recognise how communicators verbally modulate their commitments when providing them information?

Previous research has shown that the basic understanding of commitments develops from around 3-years of age, and it has been suggested that at this age children start to appreciate the normative consequences that collaborative activities entail (Tomasello, 2009, 2018, 2020). This normative turn would occur when children are around three years old, and is documented in many experiments (e.g., Rakoczy et al., 2008; Schmidt et al., 2016; Vaish et al., 2011; Warneken & Tomasello, 2013; Wyman et al., 2009). An alternative hypothesis for such sensitivity is that around this age children start to restrict the scope of commitments, thus recognising that commitments are binding only under specific circumstances, and not any time that it would be favourable for them (Michael & Székely, 2018). Younger children in fact were found to complain and to re-engage with their partner after their stoppage, or to help to complete their partner's task when this was accidentally interrupted (Green et al., 2021; Warneken et al., 2012). Either way, the literature suggests that three-year-olds possess an understanding of the normative implications of their own commitments and are willing to pay costs in order to respect them (Gräfenhain et al., 2009, 2013; Hamann et al., 2012). In fact, although younger children complain after a partner's commitment violation, only older children hold their partners responsible for intentional but not for unintentional commitment violations such as the ones due to accidental events or to partner's inability (Kachel et al., 2018). Children seem to be also sensitive to 'breaking the commitments the right way' (see Chennells & Michael, 2022): they were more inclined to 'apologise' or to acknowledge an upcoming commitment failure more often when an agreement was in place (Gräfenhain et al., 2009), and they were also found to be less resentful towards their partners who failed to fulfil a previous commitment when the partner had excused herself or asked permission beforehand (Kachel et al. 2019).

Beyond children's sensitivity to the normative aspects of commitment in collaborative contexts, there is also evidence that, already from a young age, children are sensitive to certain cues of commitment in discourses. However, findings are conflicting, and they show that only later in development children can modulate their trust in view of these cues and they manifest also a tendency to evaluate the sources on the basis of such cues.

Children between 6 and 9-year-olds were found to favour claims to first-hand evidence over claims to second-hand evidence, though this preference is more evident in older children (Aboody et al., 2022; Fitneva, 2008; Lane et al., 2018; Ozturk & Papafragou, 2016). Beyond recognising that first-hand evidence claims cue speaker commitment, i.e., they increase statement believability, children were found to judge second-hand evidence claims as less committal (Robinson & Whitcombe, 2003). Children seem to be able not only to scrutinise the validity of the arguments themselves, but to use this information to evaluate its source and modulate selective learning. 4-year-olds selectively trusted previously accurate rather than inaccurate informants (Koenig et al., 2004; Koenig & Harris, 2005). Furthermore, 4-to-8-year-olds preferred to reward (and selectively shared relevant information with) accurate informants over inaccurate informants (Li & Koenig, 2020; Ronfard et al., 2019).

In the following two chapters, I will present two studies that investigate how children react to commitment violations in these two different settings (cooperative and communicative). Using a protest paradigm, Study 5 investigates whether 3-year-olds tend to protest less when a puppet defect a joint commitment (i.e., abandon a joint activity) if the puppet faces a conflicting moral dilemma such as helping another agent in distress). Study 6 investigates the effect of different source claims on how much 6-to-7 years-old children believe a given assertion and how accountable they hold the speaker for the truth of that assertion.

Chapter 5. Three-year-olds' reactions to violations of commitments to joint goals

Most of human social life is based on coordinating with social partners to achieve greater goals than we could achieve alone, but also to build reciprocal trust, cultivate relationships, and depend on each other. Any social activity requires us to predict and to adjust to our partners' behaviours, but such predictions are difficult when partners face potential conflicting motivations. Commitments are particularly useful tools, because they reduce the uncertainty about others' behaviour by stabilizing fluctuating motivations and helping to rely on one another (Michael & Pacherie, 2015). Commitments also ground normative obligations, which entitle individuals to reproach partners who fail to maintain their commitment (Darwall, 2006; Gilbert, 2014; Tomasello, 2020). Feeling committed and understanding when partners are committed are thus key skills for navigating the social world (Gilbert, 2014, 2017; Michael et al., 2016a; Searle, 2010; Tomasello, 2009, 2016).

However, understanding when it is appropriate to release partners from their commitments is also an important part of understanding what commitments entail. In fact, every time that a commitment is in place, people have an implicit expectation about its scope and priority. When is it the case that someone is expected *to be released from* their commitments? Philosophical analyses have suggested that we should release our partner from any obligation anytime the commitment conflicts with fulfilling a moral obligation (Shpall, 2014). Imagine, for example, Sarah and Melissa hiking together on a Sunday afternoon. If Sarah suddenly leaves to join other friends in a bar downtown, Melissa is likely to be annoyed and to demand an explanation or an apology from Sarah. However, if Sarah as medical doctor leaves to assists a man who passes out, it would be strange of Melissa to demand from Sarah to follow through on her hiking commitment instead. In both cases, Sarah intentionally breaks her commitment to hike with Melissa and leaves; however, she had different motives for doing so. In the first case, her commitment conflicts with a selfish motive to have more fun with other friends. In the second case, the commitment conflicts with her moral obligation towards the patient. It has been argued that in the second case the moral obligation should outweigh the previous commitment (Shpall, 2014), thus the commitment should not be fulfilled.

Based on these arguments, we empirically investigated whether participants release their partner from an obligation to follow through when the commitment conflicts with a moral duty. To our knowledge, this idea has never been tested empirically, either in adults or children. We designed a novel cooperative game that required two partners to coordinate their actions to obtain rewards. The commitment was established by a repeated joint activity, in which children and their partner were interdependent and coordinated (Gilbert, 1990; Roberts, 2005). In the

selfish motive condition, the partner left the game to play another game; in contrast, in the moral motive condition, the partner left to help another individual. We predicted that participants would release their partner from a commitment more frequently in the moral motive condition compared to the selfish motive condition.

Previous research has shown that the basic understanding of commitments develops from 3 years of age, when children start to form joint goals when engaged in joint activities (Gräfenhain et al., 2009, 2013; Hamann et al., 2012). Three-year-olds also distinguish between situations in which their partner breaks their commitment intentionally (for selfish motives), or for unintentional reasons as accidents or inability (Kachel et al., 2018). These findings showed children's discernment of intentional and non-intentional commitment breakings; however, it is unclear whether partners' intentional commitment breakings are evaluated considering their motives. Two recent studies investigated whether children distinguish between situations in which they themselves and a third agent faced either selfish or moral motives to break an explicit commitment (i.e., a promise). The results showed a ceiling effect, with both 3- and 5-year-olds equally committed to keep their own promise no matter the motive for breaking it; however, when asked to choose the ending of a story involving a third-parties social interaction, older children preferred others' promise breaking more when they had moral than selfish motives, while younger children showed the opposite preference (Kanngiesser et al., 2021). There are several possibilities that explain this finding. While their own commitment to the experimenter was perceived as overly binding, and they may not have been able to report a relevant justification for breaking their promise, choosing a 'selfish' promise breaking ending may reflect 3-year-olds' prediction rather than their normative preference.

In contrast to Kanngiesser et al. (2021), who focused on children's motivation to keep their own commitment and preference for third-party behaviour, we tested children's reaction to their partner's commitment breaking. In the current study, our main aim was to investigate whether children are sensitive to the motives leading their own partners to intentionally break commitments, i.e., when children are themselves the ones suffering the consequences of the violation. We predicted that children are more likely to release their partner from the commitment when this conflicts with a moral duty, compared to a situation in which this conflicts with a selfish motive.

Our secondary aim was to investigate to what extent children's sophistication to understand the scope of commitments is related to their cognitive skills, namely understanding others' mental states and justificatory abilities. It has been suggested that children's theory of mind skills are related to their appreciation of commitments–specifically, their ability to

understand agents' motives when undertaking a commitment (Mant & Perner, 1988). We reasoned that those participants who satisfactorily recognized others' mental states will be better able to assess the appropriateness of the partner's motive in the moral condition. Therefore, we predicted that that theory of mind skills will predict an increased tendency to release the partner in the moral condition. Additionally, violations of moral expectations are often followed by persuasive attempts to justify one's actions (Mercier & Sperber, 2011); previous research has shown 3-year-olds' preference for non-circular arguments, suggestive of an appreciation of what counts as good justification (see e.g., Corriveau & Kurkul, 2014; Koenig, 2012; Mercier et al., 2014). We reasoned that children who satisfactorily justify their choices would also be able to properly evaluate their partner's justification. Therefore, we predicted that justification skills will predict an increased tendency to release the partner in the moral condition.

Methods

Participants

Participants were 60 3-year-old children (*M* = 3.44, *SD* = 0.28; 32 females), recruited from a database maintained at the Wa.R.Ks Group and from a private nursery in the University of Warwick. Participants came from families of heterogeneous socioeconomic background, and 3.3 % had a racial background other than White (i.e., Black). Additional children were tested but were excluded from the sample because they failed a pre-test task (*N* = 7), or were unable or unwilling to complete the trial (*N* = 8). Prior to the study, parents had given written informed consent for their children to participant in the study.

Children were tested individually in a quiet room in the Psychology Lab or in the Warwick University Nursery between April 2019 and March 2020. Each session lasted approximately 25 minutes. The experiment was conducted in accordance with the Declaration of Helsinki; all procedures were approved by the Humanities & Social Sciences Research Ethics Sub-committee (HSSREC) at Warwick University.

The study was pre-registered on OSF.org, with the sample size, coding scheme, planned analyses and participant exclusion criteria specified. The pre-registration document is available at ***https://osf.io/xfr87/***.

Materials and Design

In a between-subjects design, each child was assigned to one of the two experimental conditions (selfish motive; moral motive), and each child received two test trials. Before playing

the main game, children received a warm-up, followed by a theory of mind task and a justification task.

The materials for the main game included two puppets (the partner puppet and the distracting puppet), sixteen stickers to be collected, two sticker charts upon which stickers could be placed, and the 'sticker box'. The sticker box consisted of a cardboard box (50 x 50 x 60 cm), with a transparent lid on top and eight tubes graspable from each side of the box. Five tubes were lined up in the box, while three additional tubes were out of reach and were used to re-bait the box for the second trial. Additionally, the sticker box was accompanied by two cardboard barriers (around 60 x 20 cm) blocking the two players to access the other side of the sticker box. Children could access the rewards (stickers) placed inside the box by collaborating with a partner. Children's partner was played by a puppet (the partner puppet, controlled by an assistant). The stickers were placed in two little transparent boxes (one for the child, and one for the puppet) attached to each tube. The tubes could be moved forward only if both the child and the partner puppet pushed them forward simultaneously. If the child and the partner puppet pushed the tube forward toward the other edge of the box, the two little boxes popped out from two little windows and the stickers were free to be collected (see Figure 5.1). Given the size of the apparatus, children were not able to operate the tube alone; without the help of the partner puppet, the tube would get stuck.

Additionally, the materials for the warm-up included coloured wooden blocks and twelve plastic toy animals (four tigers, four horses, and four sharks) and three plastic sheets (featuring a jungle, a farm, and a sea). These three plastic sheets and twelve toy animals (a toy tiger, horse, shark, giraffe, pig, sea lion, and a toy elephant, cow, octopus, polar bear, camel, and unicorn) were used for the theory of mind task and for the justification task.

Figure 5.1. The sticker box game. The child held the tube from their side of the box. When the child and the partner puppet held the tube simultaneously, they were able to push the tube towards the other edge of the box where the two transparent boxes (highlighted in red in the picture) passed through the windows (highlighted in blue in the picture), and they were able to access the stickers.

Procedure

Warm-up phase. After entering the test room, the experimenter (E) introduced the child to the two puppets (the partner puppet Jaimie, and the distracting puppet Alex). The distracting puppet left the room after a short conversation, while the partner puppet and E stayed in the room and played unrelated warm-up games. Sometimes the partner puppet made mistakes (e.g., putting a red block on a blue tower, or placing a plastic horse toy on a plastic sheet representing a jungle) and the child was encouraged to verbally correct her when this happened. This served as a pre-test to ensure that children were comfortable talking with the partner puppet. The child had four chances to correct the partner puppet at least once and pass the pre-test.

Theory of mind task phase. The theory of mind task consisted in evaluating six toy animals' mental states in light of their previously stated desires. The task was adapted from Rakoczy, Warneken & Tomasello (2007) and served to measure participants' ability to recognize others' mental states (specifically, participants' engagement in desire reasoning), which is a skill generally assumed to be linked to the capacity to recognize commitments (Mant & Perner, 1988). E placed three plastic sheets (depicting a jungle, a farm and a sea) in front of the child

and presented a toy tiger, horse and shark which stated their desire to go to their preferred location (e.g., the tiger stated '*I want to go the jungle'*). Then, E placed all three animals in one fixed location (i.e., the sea), and asked the child first where did each animal wanted to go ('*Where did the (animal) want to go? The jungle, the farm, or the sea?*'), and then whether the animal was happy or sad (*'Is the (animal) happy or sad now?'*). The task was then repeated once with a second set of animals (a toy giraffe, pig and sea lion, with all being placed in the jungle).

Justification task phase. The task consisted in choosing a location for six toy animals (a toy elephant, cow, octopus, camel, polar bear and unicorn) and providing a justification to the partner puppet about the choice. The task was adapted from Domberg, Köymen, and Tomasello (2018) and served to measure participants' ability to provide justifications for their previous choices. We reasoned that this capacity might be linked to an understanding of the type of motives that justify a partner to break a commitment. E presented the child with each animal, and the partner puppet asked the child to put the animal in one of the three locations depicting a jungle, a farm and a sea ('*Where do you want to put the polar bear? In the jungle, the farm or the sea?')*. The partner puppet then asked the child to justify the choice, e.g., *'Why did you put the polar bear in the sea?*'.

Demonstration and training phase. E revealed then the main game (a sticker box) and taught both the child and the partner puppet how to play it. The child and the partner puppet sat by their respective sides of the sticker box, and E instructed them that they needed to push the upper tube forward together. Once the tube is pushed to the other edge of the sticker box, two little transparent boxes containing stickers popped out of the sticker box (see Figure 1). The first tube was pushed (and the first two stickers collected) under the supervision of E, who ensured that the child understood the interdependent aspect of the game, namely that it would not be possible to collect the sticker if the partner puppet and the child did not push the tube together. In order to discourage the child from moving around the sticker box or leaving their position at their side of the box, E placed two cardboard barriers and stated that both the partner puppet and the child could not cross it. E showed the child and the partner puppet a sticker chart, with eight sticker placeholders, that they could use to stick the collected stickers, and left the room.

Main game phase (2 trials)

Joint activity phase. To create a joint commitment, the partner puppet simply started pushing the tube on their own but failed to move it. In case the child did not act, the partner puppet would complain about not getting any sticker. In no circumstance did the child and the partner puppet verbally agree to play together, nor did the partner puppet explicitly request

that the child do so. The child and the partner puppet each collected three stickers together (by pushing three of the four remaining tubes). During the game, the partner puppet cheerfully commented on the progress (‘*Up the mountain, almost there!, down the mountain’*) and showed a sticker she collected to the child, encouraging the child to do the same (‘*Yeeeh! Look! I got a blue alien! What did you get?’*). When they have started to push the last tube to collect the last sticker, the distracting puppet opened the door and stood at the entrance of the room.

Manipulation phase. In the selfish motive condition, the distracting puppet interrupted the game and lured the partner puppet to play another game in another room (‘*There are many colourful balls here!*’) In the moral motive condition, the distracting puppet interrupted the game by expressing the need for assistance with a small injury (‘*My finger is hurting!’*). The script played out as followed:

- Distracting Puppet: ‘Hey Jaime! There are many colourful balls here!’ [or ‘Hey Jaime! My finger is hurting’]
- Partner Puppet: ‘There are colourful balls?’ [or ‘Your finger is hurting?’]
- Distracting Puppet: ‘Yes, there are balls for puppets! I am going to play with many balls now’ [or ‘Yes, my finger is hurting! I need a puppet to put on a plaster!’]
- Partner Puppet: ‘You’re going to play with many balls? Oh, I have a ball’ [or ‘You need a puppet to put on a plaster? Oh, I have a plaster’]
- Distracting Puppet: ‘I am going to the room next door’.

In both conditions, the distracting puppet stated that she is going to the room next door and then left the test room. The time-intervals in the test phase were measured via a watch worn by the partner puppet.

Test phase – open-ended part. The partner puppet interrupted the sticker game and alternated looking between the child and the door, and after 2 s it started approaching the door. After 2 s the partner puppet addressed the child, repeating the motive for leaving (‘*Alex is in the room next door! And there are many colourful balls there/and he needs a puppet to put on a plaster on his finger!’*), and then walks towards the door. After 5 more s, the partner puppet again addressed the child asking what to do (*‘What shall I do?’*).

Test phase – forced-choice part. After additional 5 s the partner puppet asked the child whether she should stay in the room or go out to the distracting puppet (‘*Should I stay here, or should I go to Alex?*’).

Then, the partner puppet left the room, while E entered again, re-baited the apparatus and left. The partner puppet entered, and played the main game with the child one more time.

Debrief phase. After having played the main game phase a second time, E and the child's caregiver came back to the room and collected with the child all the remaining stickers. E thanked children for playing and gave them a certificate of achievement, the collected stickers, and a small toy.

Coding and Reliability

All the sessions were videotaped. Children's responses in theory of mind and justification tasks, as well as children's reactions to the partner puppet's failure to finish collecting the stickers were transcribed by a native speaker. Transcripts were then coded by the first author unaware of the condition.

Release score. The main measure was whether the child denied or granted the partner puppet release from their commitment during the test phase. In each part of the test phase (open-ended part and forced choice part), children were assigned a score from -0.5 to 0.5 according to whether they granted or denied release to the partner puppet. If children granted release, and if they did not explicitly release the partner puppet but manifested signs of release (that is, references to the possibility to follow the distracting puppet to get involved in the alternative activity), they were assigned a score of 0.5. If children instead explicitly protested (i.e., denied release), and if they did not explicitly protest, but manifested signs of protest (i.e., references to the main game), they were assigned a score of -0.5. If children were silent or did state something unrelated to both the main game and the alternative option, they were assigned a score of 0 (see Table 5.1 for examples). As each test phase included a part in which children could express freely their mind (open-ended part) and a part in which they were invited to make a choice between two pre-determined choices (forced-choice part), the release score was the results of the sum of the scores of each part. This resulted in a 5-point release score scale (range -1; 1) indicating the level of release that children manifested during the test phase: a highest score of 1 indicates a consistent expression of release in the two parts, whereas a lowest score of -1 indicates a consistent expression of protest in the two parts of the test phase.

Table 5.1. *Examples of utterances in the open-ended and forced choice parts.*

Label	Description	Example	Release Score
Full protest	Utterances directed to the partner with reference to main game and involving a normative dimension – with the occurrence of terms such as: must, ought, should, may, right/wrong, good/bad, have to.	'You should help me take the yellow one out' 'You should actually stay here' 'We should do this one'	-0.5
	Utterances aimed to re-engage the partner with the main game or indicating that the participant wants the partner to stay and play (including imperatives).	'Let's get some more stickers!' 'Help me!' 'Stay here'	
Signs of protest	Interjections.	'No!'	
	Utterances directed to the partner aimed to direct the partner's attention to the fact that the main game is incomplete.	'There's one more' 'I am not finished yet' 'I can't do this'	
Sign of release	Utterances directed to the partner that refer to the alternative option, including statements aimed to direct the puppet's attention to the alternative option.	'Seems he needs some help' 'Where is he gone?' 'Has he hurt his finger?'	0.5
	Statements that make explicit that the child agrees with the decision of leaving.	'Oh, OK'	
Full release	Utterances directed to the partner that refer to the alternative option and involve a normative dimension (with the occurrence of terms such as: must, ought, should, may, right/wrong, good/bad, have to).	'You should go to Alex' 'You need to go' 'You have to give the balls to your friend'	
	Statements aimed to direct the puppet to take up the alternative option, including imperatives.	'Alex' 'Go to Alex' 'Help him!'	

To establish reliability, a naïve coder blind to the conditions and the hypotheses of the study coded the whole data sample. The two coders were in almost perfect agreement (*Cohen's k* = 0.96). Disagreements were solved by discussion between the authors, unaware of the conditions.

Theory of mind and justification scores. All children's responses to the six trials of the ToM task were coded as incorrect (0) or correct (1); e.g., if the tiger was placed in the sea and the child responded '*the tiger is sad'*, this response was coded as 'correct', provided that the trial was valid, i.e., the child correctly recalled the animals' previously expressed desires (e.g., if the child responded '*the tiger wanted to go to the farm*', thus wrongly recalling the animal's desire, their response was discarded). Children were assigned an average ToM score (0 – 1) based on the amount of correct answers divided by the total amount of valid trials.

All children's responses to the partner puppet's question why children put the animal in the chosen location were scored as relevant (1) or irrelevant (0). Relevant justifications were arguments based on the coding scheme from Domberg et al. (2018). Irrelevant justifications included circular arguments (e.g., '*because I did it'*), references to the child's desire (e.g., '*because I want so'*) and no responses. Children were assigned an average justification score based on the amount of relevant justifications provided in six trials (0-1).

Again, to establish reliability a naïve coder blind to the conditions and the hypotheses of the study coded the whole sample. The two coders were in almost perfect agreement with the ToM scores (*Cohen's k* = 0.91) and the justification scores (*Cohen's k* = 0.93). Disagreements were solved by discussion between the authors, unaware of the conditions.

Data Analysis

We analysed children's reactions by running three separate cumulative link mixed models to assess the effect of the experimental condition on participants' release scores, and the secondary measures on the release scores from participants who participated in the moral condition. For each subject, we included both trials in the analyses.

The main model included condition, trial number and gender as fixed effects, and participant ID as a random effect. The secondary models included ToM score or justification score, trial number and gender as fixed effects, and participant ID as a random effect. The null models included trial number and gender as fixed effects, and participant ID as a random effect. Model comparison was done using likelihood ratio tests (Dobson & Barnett, 2008; Forstmeier & Schielzeth, 2011). Statistics were done using R version 3.6.3 (R Core Team, 2020) and the

packages 'ordinal' (Christensen, 2019), 'lme4' (Bates et al., 2015) and 'afex' (Singmann et al., 2021).

This final analysis presented in the manuscript deviates from the pre-registered version insofar as we decided to build three cumulative link mixed models, which better accommodate ordinal data compared to linear models (Liddell & Kruschke, 2018). (See ***https://osf.io/bhkz8/*** for more details).

Results

Participants tended to protest more than to release their partner in both conditions (see Table 5.2). With regards to their performances in the secondary tasks, participants presented 63% correct answers in the ToM task (M = 0.633, SD = 0.301) and 50% relevant justifications in the justification task (M = 0.503, SD = 0.408).

Table 5.2. *Full distribution of reactions from the participants in the test phase (N = 60, 2 trials, open-ended + forced choice parts).*

Condition	Full protest	Signs of protest	No Reactions	Signs of release	Full release
Total	91/240	33/240	38/240	13/240	65/240
Moral motive	43/120	14/120	20/120	10/120	33/120
Selfish motive	48/120	19/120	18/120	3/120	32/120

Main Model

We analyzed participants' release scores in the two conditions. We compared our model with a null model using a likelihood ratio test, but the model did not fit the data better than the null model, $\chi^2(1) = 0.58$, $p = .448$. (See ***https://osf.io/bhkz8/*** for more details). There was thus no significant effect of condition on children's release scores (see Figure 5.2).

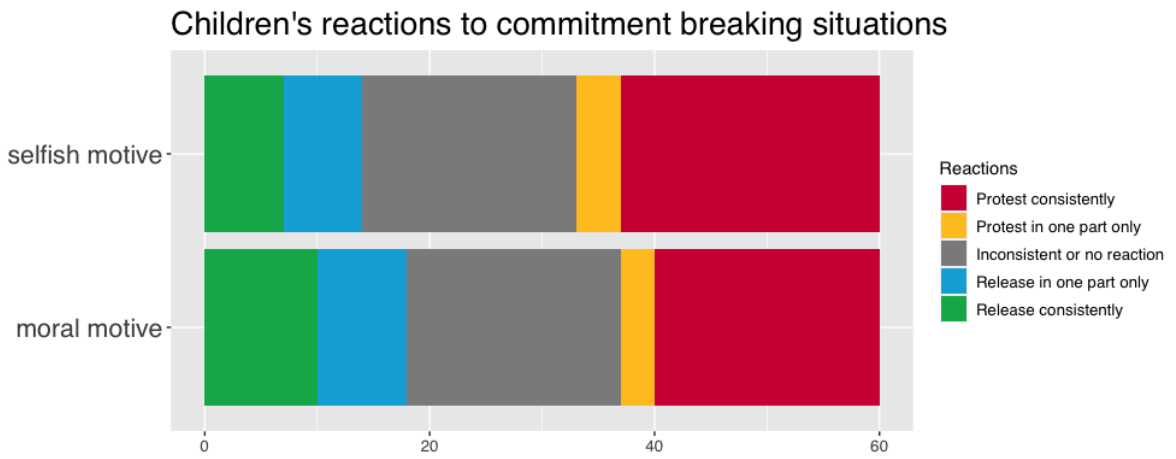

Figure 5.2. Children's release scores in the two conditions. Since each participant participated in two test trials, the figure shows up to two scores per participant (one per trial). The score is a 5-point scale, and for better readability we reminded in the legend how the child is expected to behave in the two parts of the test phase (open-ended and forced choice). Overall, children protested more (42%) than released (27%).

Secondary Models

We analysed participants' release scores in view of their ToM scores in the moral condition only. We compared our model with a null model using a likelihood ratio test, but the model did not fit the data better than the null model, $\chi^2(1) = 0.05$, $p = .817$. There was thus no significant effect of ToM score on children's release scores.

We analysed whether participants' release scores in view of their justification scores in the moral condition only. We compared the model with a null model, and the model was a significantly better fit compared to the null model, $\chi^2(1) = 7.35$, $p = .007$. The model showed a significant effect of the justification score on the release scores, *estimate* ± *SE* = -2.42, 0.94, *95% CI* [-4.25, -0.58] $\chi^2(1) = 7.35$, $p = .007$ (see ***https://osf.io/bhkz8/*** for more details). These results manifest however an unpredicted pattern: a higher justification score predicted a lower release score, indicating a tendency to release the partner puppet less often in the moral motive condition (see Figure 5.3).

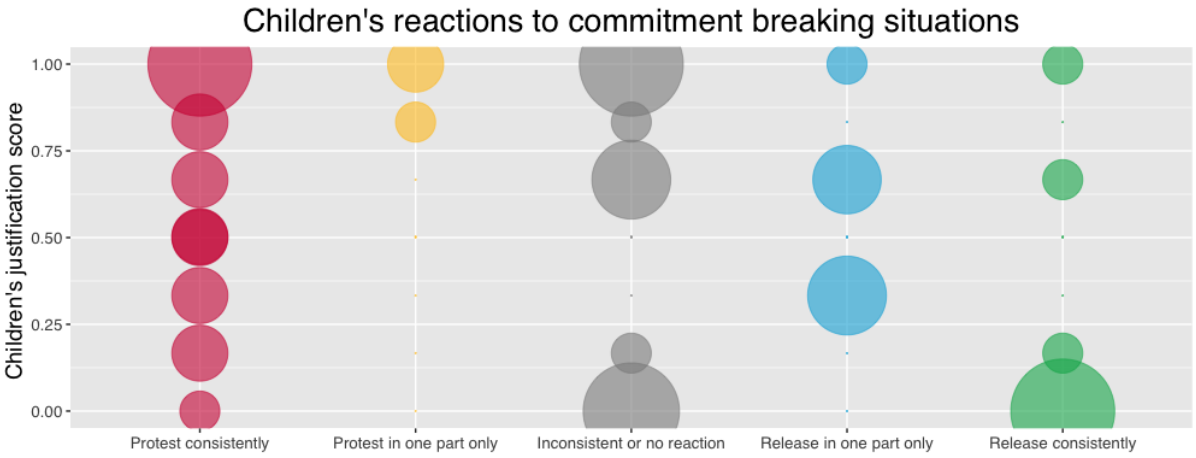

Figure 5.3. Participants' distribution of justification scores and release scores in the moral condition. We can see that children with higher justification score were more frequent to protest consistently, while children with lower justification score were more likely to consistently release the partner.

Discussion

To properly master the concept of commitment, it is crucial to understand when commitments are not supposed to be fulfilled. Our results indicate that 3-year-olds did not manifest different reactions when their partner interrupted the joint activity to do something else merely for fun (selfish motive) than to help another agent in distress (moral motive). However, they also surprisingly show that, when leaving to respond to a moral imperative, children with better argumentative skills tended to release their partners less often.

This is the first study investigating whether 3-year-old children evaluate why a partner is intentionally breaking an implicit commitment (see Table 5.3). While children at this age were found to distinguish between intentional and unintentional commitment breakings (Kachel et al., 2018), and between excused and not excused commitment breakings (Kachel et al., 2019), our results show that 3-year-olds' evaluation of the scope of commitment is not yet fully developed. This would be consistent with recent findings showing that young children are reluctant to break their own promises, irrespective of the motive to break them, but also that at this age children judge it to be more permissible for an agent to break a promise for selfish than for moral motives (Kanngiesser et al., 2021). This is also the first study measuring children's tendency to actively freeing their partner from their previous obligation. Previous literature has focused on detecting children's signs of protest, overlooking those signs of release that would indicate a finer appreciation of the scope and priority of the established commitment. In our study, children overall protested more often (42%) than they released their partner (27%): this suggests that most children still overestimate the scope of commitment

(Michael & Székely, 2018), but also shows that children take initiatives to release their partners, despite releasing requires a more complex assessment of the situation.

***Table 5.3**. Review of studies investigating children's understanding of intentional commitment breakings.*

Study	Role in the commitment	Implicit or explicit commitment	Measure	Finding
Kachel et al., 2018	2nd person	Explicit	Protest, tattling, teaching	3-year-olds react differently to intentional and non-intentional commitment breakings.
Kachel et al., 2019	2nd person	Explicit	Protest	3-year-olds react differently to excused and non-excused commitment breakings.
Kachel & Tomasello, 2019	1st person	Both	Bribe accept	5-year-olds resist bribes even when commitment is implicit, though to a lesser degree.
Kanngiesser et al., 2021 – Study 1	1st person	Explicit	Keep promise	3- and 5-year-olds keep their promises no matter the motive to break it.
Kanngiesser et al., 2021 – Study 2	3rd person	Explicit	Ending choice	5- but not 3-year-olds prefer situations when agents break their promise for moral motive.
Current Study	2nd person	Implicit	Protest-release	3-year-olds do not react differently to moral and selfish commitment breakings.

That this sensitivity is not yet fully developed is shown also by the unexpected relation between children's responses and the justification scores. Our data show that children who were better able to justify their own choice (i.e., had higher justification scores) were also

surprisingly less inclined to release their partner in the moral condition. This finding is intriguing: we may speculate that the children who were better able to justify their own choice were also more at ease in entering into an argument with a puppet who was manifesting the intention to abandon the joint activity.

It is worth noting that some children used normative language both when they released and when they protested. This is particularly striking given that our paradigm did not involve any verbal commitment or agreement to play, and the experimenters never used normative nor we-language, in contrast to most studies in the literature (with some exceptions; see e.g., Kachel & Tomasello, 2019). This finding is consistent with the idea that commitment should be considered a 'graded' phenomenon, and not a binary one, and as such does not necessarily need to be verbally explicit to have some normative power (Michael et al., 2016a; Tomasello, 2020).

In addition, we aimed to minimize the potential impact of the experimenter as an authority figure. In most previous studies, joint commitments have been initiated by the experimenter rather than by the children spontaneously (e.g., Gräfenhain et al., 2009; Hamann et al., 2012). Thus, it is possible that in some instances children respond strongly to certain factors because of an underlying commitment to the experimenter, an authority figure whom the children asymmetrically relate to and who may have different expectations than what an equal partner would have (A. P. Fiske, 1992). In our study the experimenter never encouraged the child to continue to play the game. Instead, the establishment of the commitment was driven by their interdependence and coordination the joint activity, suggesting that this aspect is decisive in building motivations and expectations for cooperation (Roberts, 2005; Gilbert, 1990; see also Bonalumi et al., 2019; Guala & Mittone, 2010; Grueneisen & Tomasello, 2020; Koomen et al., 2020; McEllin et al., 2022; Michael et al., 2016b; Rusch & Lütge, 2016).

Our study raises additional new questions for further research. For example, does the tendency to release partners develop differently from the mere tendency to inhibit protesting? Do children manifest additional implicit expectations about the scope of commitments? Would these expectations differ across multiple cultures (or sub-cultures), in particular across non-WEIRD children (Henrich et al., 2010; Nielsen et al., 2017)? Also, given that protests can be considered strategies of partner control (Tomasello, 2020), it would be interesting to assess whether a partner's tendency to release has an impact on children's partner choice strategies (Barclay, 2017; Baumard et al., 2013; see also Isella et al., 2019).

In summary, our results extend prior research on the early development of an understanding of commitment and suggest that 3-year-old children do not evaluate the

motives why a partner intentionally breaks an implicit commitment. Our paradigm allows us nonetheless to observe that children's justificatory abilities predict their tendency (not) to release their partner. This finding suggests that the capacity to evaluate a partner's motives to break a commitment is emerging but is not yet developed.

Chapter 6. Six-to-seven years-olds' reactions to violations of commitments to assertions

Most of humans' social life is made possible by the fact that people communicate. On the one hand, communication allows people to acquire relevant information about their (cultural) environment that would not be acquirable otherwise (see Csibra & Gergely, 2009; Tomasello, 1999). The possibility for communicators to influence others' minds and behaviours gives them the power to exploit communication, potentially deceiving their listeners whenever incentives to do so are present. For communication to be evolutionary stable, listeners must be able to counteract speakers' communicative power. In fact, listeners need to be (and they are) able to gauge whether the information transmitted is reliable or not, by means of assessing whether its content is credible or whether its source is benevolent and/or competent. Such assessment enables listeners to protect themselves against deception and eventually to reject information that is deemed to be misleading. This set of capacities to scrutinise communicated information was labelled 'epistemic vigilance' (Sperber et al., 2010). Deceiving communication and epistemic vigilance, then, are competing forces that allow both communicators to influence others and listeners not to be at the mercy of communicators' incentives to provide misleading instead of relevant information.

One key element in the arms' race between deceiving communication and epistemic vigilance is commitment. In fact, communicators can increase the chances that their messages are accepted is by way of modulating their commitment to the message conveyed. The more communicators signal their assurance that their message is reliable, the more they signal their willingness to pay some costs if their message turns out to be unreliable. In other words, commitments are credible because they entail putting communicators' own reputation at stake (cf. also Brandom, 1994; Geurts, 2019; Heintz & Scott-Phillips, in press; Nesse, 2001). For example, if a speaker asserts that the train is leaving at 7.02 pm, they are not only conveying the message that the train is leaving at 7.02 pm: they are committing to the truth of their message, and they are thus putting themselves in the position of being negatively judged (or punished) if it turns out that the train does not leave at 7.02 pm. Suppose that the speaker not only asserts that the train leaves at 7.02 pm, but they assert for example that they are sure that the train leaves at 7.02 pm; or while asserting that the train leaves at 7.02 pm it is mutually manifest that the listeners have a tight connection to catch that train and may lose the train if it leaves only few minutes earlier (see also Van Der Henst et al., 2002). In these cases, the information that the train is leaving at 7.02 pm (and not at 7.01 pm) is more credible, and the speaker would consequently be even more vulnerable to the listeners' reproach if this turns

out not to be the case[22]. By tracking speaker commitment and its strength, listeners can gauge to what extent they should accept a claim based on the assurance of the speaker.

Communicators and speakers can hammer the credibility of their conveyed message by means of referring to their epistemic authority over the truth or relevance of the conveyed message. In fact, the source of one's claim (e.g., marked by evidentials) is evaluated as differentially committing speakers to the truth or relevance of their message: speakers who claim that the source of their information is their own first-hand experience (e.g., 'I saw that…') are believed more and hold more accountable if the information is false compared to speakers who claim that the source of their information is second-hand evidence (e.g., 'I was told that…) (Mahr & Csibra, 2021). Various other phenomena have been shown to similarly modulate speaker commitment, i.e., statement believability and speaker's accountability. Expressions of confidence increase the believability of a message, but they also lead to higher direct and reputational costs if the message is found to be unreliable (Vullioud et al., 2017). The use of modal expressions is instead taken to be a cue of speakers' willingness to distance themselves from their message, thus decreasing its believability (Degen et al., 2019). Pragmatic implicitness is equally interpreted as a pragmatic cue of decreasing the commitment to the message (Mazzarella et al., 2018)—although the distancing effect provided by the use of implicitness meets its limitations, for instance when the implicit meaning is very strongly implied, and the listeners is relying on the truth of the message (Bonalumi et al., 2020, 2021).

There is currently some tentative evidence that, already from a young age, children are sensitive to certain cues of speaker commitment: they modulate their trust in view of these cues and they manifest a capacity to scrutinise the validity of speakers' arguments and justifications. It is well established that, already from a young age, children can discriminate between 'good' arguments and 'bad' arguments such as circular arguments. Children as young as 2 and 3 manifested a preference for non-circular over circular arguments (Castelain et al., 2018; Mercier et al., 2014). Japanese pre-schoolers showed a similar preference (Mercier et al., 2017), and Maya pre-schoolers were found to preferably believe assertions supported by claims supported by first-hand evidence (e.g., 'I saw that…') over claims supported by circular arguments, even when given by dominant informants (Castelain et al., 2016). Koenig (2012) found that both 3- and 5-year-olds recognized that an informant had a 'better way of thinking' when their claim was supported by first-hand evidence, deduction, or testimony rather than when their claim was justified by a 'bad' argument (desiring, pretending or guessing that what

[22] As stated in the Introduction, this is the case with assertions, and even more so with promises. But as Geurts (2019) pointed out, all speech acts entail a series of discursive and not discursive commitments.

is claimed is true). However, children did not show any significant preference for informants whose claim was supported by first-hand over second-hand evidence, suggesting that second-hand evidence claims are considered as credible. On the other hand, other studies reported such preference in older children. Bulgarian 6-to-9-year-olds expected a third party to believe statements marked by first-hand evidentials compared with statements marked by second-hand evidentials[23], though this preference is more evident in older children (Fitneva, 2008); Turkish children showed a similar pattern (Ozturk & Papafragou, 2016; see also Çelik et al., 2022). Lane and colleagues (2018) similarly found that 6-to-8-year-olds, when judging improbable events, favoured claims to first-hand evidence over claims to second-hand evidence. Beyond recognizing that first-hand evidence claims cue speaker commitment, i.e., they increase statement believability, children were found to judge second-hand evidence claims as less committal. 3-year-olds (but not 2-year-olds; see Mascaro & Kovács, 2022) already took second-hand evidence to be less reliable than direct evidence (i.e., their own experience), unless there were reasons to believe that those who provided second-hand evidence were better informed (Robinson & Whitcombe, 2003). While pre-schoolers still believed hearsay (second-hand evidence) more than unjustified claims, older children showed more scepticism and believed hearsay less than unjustified claims (Lane et al., 2018; Danovitch & Lane, 2020).

Children seem to be able not only to scrutinise the validity of the arguments themselves, but to use this information to evaluate its source and modulate selective learning. Both 3- and 5-year-olds were found not only to prefer non-circular over circular explanations, but they also took this information into account and selectively learnt from informants who previously provided non-circular explanations (Corriveau & Kurkul, 2014). Similarly, 4-year-olds selectively trusted previously accurate rather than inaccurate informants (Koenig et al., 2004; Koenig & Harris, 2005). Furthermore, 4-to-8-year-olds rewarded inaccurate informants less than informants who provided accurate information or no information at all (Ronfard et al., 2019), although this attitude is not corroborated by other studies—e.g., Li and Koenig found that 4-year-olds rewarded speakers independently of their accuracy, although they did selectively share relevant information with accurate informants more often than they did with inaccurate informants (Li & Koenig, 2020).

To sum up, there is evidence that at least 6-years-old children tend to selectively learn and reward accurate informants, as well as they tend to believe first-hand evidence claims more

[23] In Bulgarian, the second-hand evidential marker indicates that the source of knowledge is not direct, but is ambiguous with regard to whether it is inferential or reportative (i.e., testimony). A similar phenomenon in present in Turkish, with the second-hand evidential marker *-mış*.

than second-hand evidence claims. It is still unknown, however, whether they would make sense of speakers' use of source claims to hold them socially accountable when such claims are found unreliable, and to guide their selective learning on this basis.

To probe this, we ran an on-line study aimed to assess (i) whether 6-to-8-years-old children believe assertions on the basis of the source claim stated by the speaker; and once the assertion is found to be false, whether children take the use of the source claim into account when (ii) choosing whether to hold an informant accountable for having mislead their audience, and (iii) selectively trust them by requesting additional information.

Methods

The study was pre-registered on AsPredicted.org, with sample size, planned analyses and participants exclusion criteria specified. The pre-registration document is available at ***https://aspredicted.org/GXN_WPB***). The methods used in this study are in accordance with the international ethical requirements of psychological research and approved by the EPKEB (United Ethical Review Committee for Research in Psychology) in Hungary.

Participants

Participants were recruited partially through a database maintained at the BabyLab of the Central European University in Budapest, and partially through the Baby Lab social networks. All children were tested individually on-line between September and November 2021. The final sample (N = 29) was composed by 6-to-7-year-old children (13 females, M_{age} = 6.67, range = 6;0 to 7;11), all Hungarian speakers. One additional child was tested but not included in the final sample because of language comprehension problems. Parents received a voucher as compensation for their participation in the study.

Materials and Design

Each child participated in a familiarisation phase and a test phase, in that order. The whole experiment lasted approximately 15 minutes and was conducted in Hungarian.

Children were shown 2 animated clips, build up on MS Powerpoint v.16.60 and streamed via the share screen function of Zoom 5.8.0 (Zoom Video Communications, 2022). Children were sitting in front of the screen in their own house. On each of these presentations, three animals and two boxes were presented on the screen. The animals at the centre of the screen manifested interest in collecting a food item supposedly hidden in one of the two boxes. Each animal informants made contrasting claims regarding the location of the desired food item. In the critical phase, one informant's statement was supported by a first-hand evidence claim

("*Láttam, hogy…*" [I saw that…]) and the other informant's statement was supported by a second-hand evidence claim ("*Valaki azt mondta nekem, hogy…*" [I was told that…]).

Procedure

Introduction phase. At first, the child was acquainted with the experimenter and with some of the stimuli that would feature in the experiment itself. The experimenter introduced some animated characters and checked that the child was able to verbally distinguish different coloured boxes. The experiment itself started with a confirmation to continue from the child.

Familiarisation phase. The experiment began with a familiarisation phase, which was composed of three parts. The purpose of the familiarisation phase was to allow the child to familiarise with the on-line setting and the dynamic between the animated characters on the screen, and to test their sensitivity to correct and incorrect information. Participants were presented an animated story that included three animated characters: an animal searcher (a bear) who stated to be looking for a food item (honey). Two informants (an elephant and a hippo) offered their opinion in response. The position of the two informants, their voices, and the colour of the boxes were randomly assigned. A few elements were instead carefully counterbalanced: (a) the location of the food item between the Belief part and the Trust part; (b) the voices of the informants between the Familiarisation phase and the Test phase; (c) the location of the informants who spoke first between the Familiarisation phase and the Test phase.

Belief part. The bear in need of information said: "*I am hungry! I have to find my honey! Hmm…*". The two informants provided then conflicting information about the location of the honey without referring to any source. For example, the elephant stated: "*Look into the red box!*" and the hippo stated: "*Look into the yellow box!*". After both informants stated their suggestion, the bear asked the child the 'Belief question': "*Where is the honey? In your opinion, where is my honey?*".

The experimenter guided the bear according to the child's choice: the bear opened first the box indicated by the child, and then the other box. At this stage, the food item was found randomly in one of the two boxes, e.g., the red box indicated by the elephant.

Reward part. After the bear obtained their honey, children were asked which informant to reward for their help. The bear found some chocolate treat, and both informants expressed interest in eating it. The bear recalled then the information collected from the informants and whether they were accurate or not: for example, that the elephant said to look into the red box, and the honey was there, and the hippo said to look into the yellow box, but the honey

was not there. Then, the bear asked the child the Reward question: "I*n your opinion, who should get the chocolate?*".

The experimenter guided the informant indicated by the child, and the informant thus reached and ate the chocolate treat. All animals ate their food.

Trust part. After the reward part, the same animated characters stood on the screen, with two new boxes of different colours appeared. The bear expressed additional interest in finding a new food item (a raspberry) and this time asked the child the Trust question: "*Where is my raspberry? In your opinion, who shall I ask where my raspberry is?*".

The experimenter enabled the informant indicated by the child to speak. This informant stated their suggestion about the location of the raspberry, which was always in the box in the opposite location compared to where the food was found in the Belief part (for instance, if the honey was previously found on the box on the left side, the raspberry would be found in the box on the right side). The bear followed the suggestion, and then found the food. The scene concluded with the bear happily enjoying their food.

Test phase. A new animated character (a giraffe) appeared on screen, along with two new coloured boxes (see Figure 6.1). The giraffe stated to be looking for a food item (a carrot). Two new animal informants (a cow and a goat) offered their opinion in response, but this time providing a reference to their source. The sources were randomly assigned to the two informants.

Belief part. The giraffe in need of information said: "*I am hungry! I have to find my carrot! Hmm…"*. The two informants provided then conflicting information about the location of the carrot, with different source claims backing their suggestions; one informant provided a first-hand evidence claim, and the other informant provided a second-hand evidence claim. For example, the goat stated: "*Look into the green box! I saw that the carrot is in the green box! I saw it!*". The cow stated instead: "*Look into the blue box! Somebody told me that the carrot is in the blue box! Somebody told me!*". After both informants stated their suggestions, children were asked by the giraffe the Belief question: "*Where is the carrot? In your opinion, where is my carrot?"*.

As in the Familiarisation phase, the experimenter guided the bear to open first the box indicated by the child, and then the other box. In this phase, however, the food item was not found in either of the boxes.

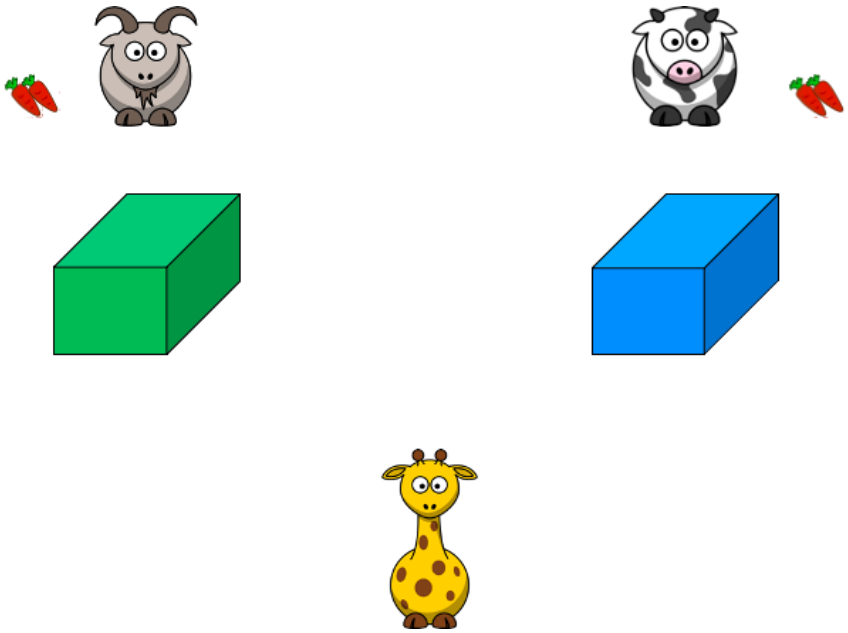

Figure 6.1. Layout of the animation during the Test phase (Belief part). The giraffe was looking for the carrot and the goat and the cow gave conflicting information: for instance, the goat informed the giraffe that the carrot is in the green box, because they saw it is in the green box, while the cow informed that the carrot is in the blue box, because somebody told that it is in the blue box.

Accountability part. Children were asked which informant should take responsibility for misleading the giraffe, and provide for the missed object. The giraffe thus expressed the intention to take a carrot from one of the two informants. Both informants expressed discomfort about the idea of having one of their carrots taken away. The giraffe recalled that the cow said that they saw that the carrot was in the green box, but the carrot was not there, and the goat said that somebody told them that the carrot was in the red box, but the carrot was not there either. The giraffe thus asked the child the Accountability question: "I*n your opinion, who should I take one carrot from?*".

Again, the experimenter guided the animal chosen by the child to have one of their carrots taken from the giraffe searcher. All animals ate their food.

Trust part. After the accountability part, the same animals stood on the screen, and two new boxes appeared. The giraffe expressed additional interest in finding a new food item (an apple) and this time asked the child: "*Where is my apple? In your opinion, who shall I ask where my apple is?*".

Again, the experimenter enabled the informant indicated by the child to speak, and the informant provided the accurate information about the location of the apple. The giraffe

followed then the suggestion given by the informant, and found the food. The scene and the experiment concluded with the giraffe celebrating and eating their apple.

Both the child and their caregiver were thanked for their time.

Coding and reliability

The videos were coded by a native speaker. To establish reliability, the main investigator coded 25 % of the whole data sample. The two coders were in perfect agreement (*Cohen's k* = 1).

Data analysis

We analysed children's responses by running a Bayesian binomial test on each dependent variable to assess the preference of choice between the two informants (correct vs incorrect informant in the familiarisation phase; first-hand vs second-hand evidence claim informant in the test phase). The analyses were performed by using R version 3.6.3 (R Core Team, 2020), and the packages 'rjags' (Plummer et al., 2021). We ran five Bayesian binomial tests in JAGS by MCMC sampling, to test the hypothesis that the proportion of children who chose the two informants is not the same. We chose an uninformative prior (with both beta parameters set at level = 1). We ran a total of 3 iteration chains, each with 3334 iterations. Following Krushke (2015), we specified a region of practical equivalence (ROPE) between 0.45-0.55 in order to estimate whether the posterior distribution credibly confirms or rejects the null hypothesis. Additionally, we estimated the Bayes Factors by using JAMOVI (The jamovi project, 2021).

The analysis was pre-registered on AsPredicted.org, specifying data collection and stopping rule (see ***https://aspredicted.org/GXN_WPB***). Given the results of the first 24 children, however we estimated that this proportion of responses would not convey more than moderate evidence for either hypothesis with *N* = 80, so we interrupted data collection. We included the data from children who were recruited beforehand (*N* = 5).

The table below summarise our predictions with regard to all the measures in both phases (familiarisation and test phase).

Table 6.1. *Summary of the predictions.*

Question	Familiarisation Phase	Test Phase
Belief question	NA	First-hand (see) > second-hand (tell)
Reward / Accountability question	Accurate > Inaccurate	First-hand (see) > second-hand (tell)
Trust question	Accurate > Inaccurate	Second-hand (tell) > first-hand (see)

Results

Familiarisation phase results

We tested the hypothesis that children would reward more often and would selectively trust more often an accurate informant than an inaccurate informant. As can be seen in Figure 6.2, most children chose to reward the accurate informant (*N* = 24) rather than the inaccurate informant (*N* = 5); however, contrary to our prediction, most children chose to selectively trust the inaccurate informant (*N* = 20) rather than the accurate informant (*N* = 9).

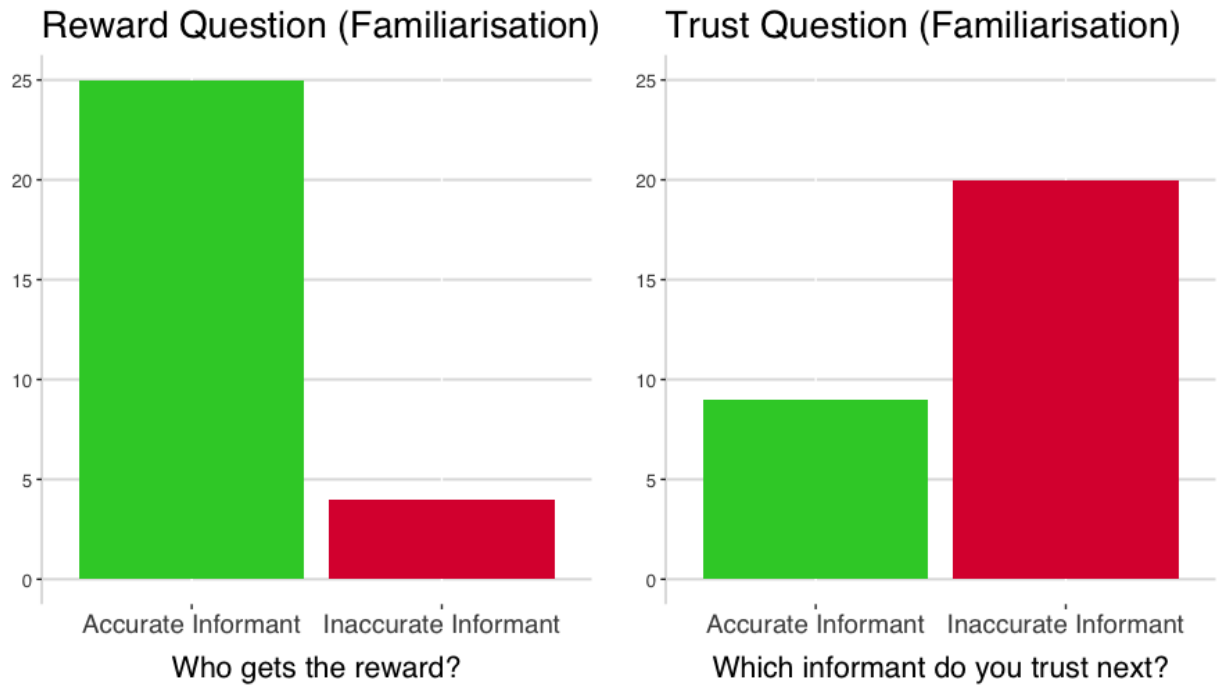

Figure 6.2. Amount of children who chose to reward and selectively trust the accurate or the inaccurate informant.

We carried out a two-sided Bayesian binomial test based on the alternative hypothesis that the proportion of children who rewarded an accurate informant was different from chance level (0.50).

As we can see from the posterior distributions plotted in Figure 6.3, the estimated median is 0.85, and the 95 % of the high density interval (HDI) [0.707 – 0.952] falls entirely outside the region of practical equivalence of 0.45-0.55 (ROPE). We set the limits of the ROPE following Kruschke (2015) This suggests that the data are better explained by the alternative hypothesis, and specifically shows that children reward accurate informants significantly more often than inaccurate informant. We additionally computed the Bayes Factor, which confirmed this finding: the computed BF was 150.69 which provides extreme evidence favouring the alternative hypothesis.

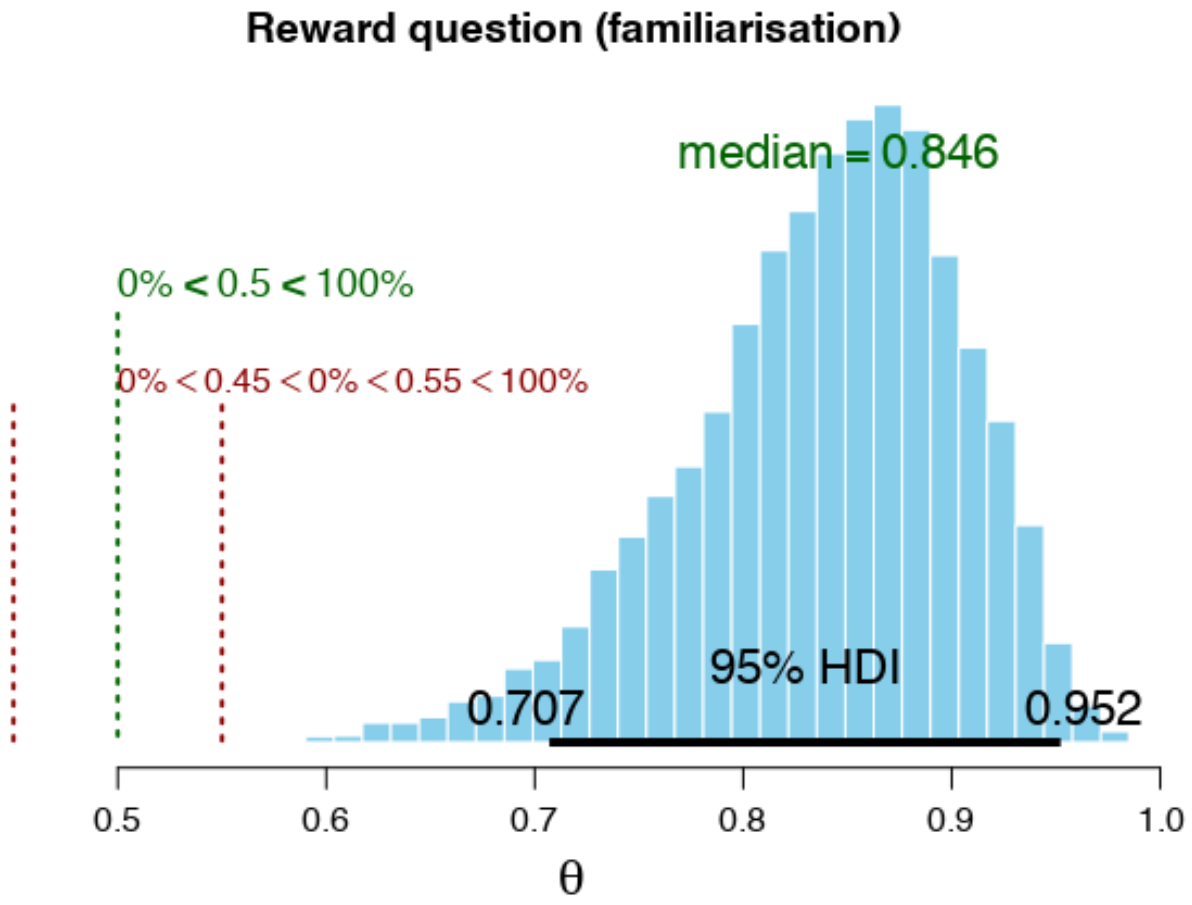

Figure 6.3. Posterior distribution showing extreme evidence for the alternative hypothesis that children rewarded an accurate informant, here coded as 1 (or an inaccurate informant; here coded as 0) significantly more often than chance (0.50).

We also carried out a two-sided Bayesian Binomial test based on the alternative hypothesis that the proportion of children who selectively trust an accurate informant was different from chance level (0.50).

As we can see from the posterior distributions plotted in Figure 6.4, the estimated median is 0.32, and the 95 % of the high density interval (HDI) [0.17 – 0.49] only partially overlaps with the ROPE (6.5 %). This suggests that we have inconclusive evidence about whether the data are better explained by the null or the alternative hypothesis. We additionally computed the Bayes Factor, which confirmed this finding: the computed BF was 1.79 which suggests high uncertainty in the model.

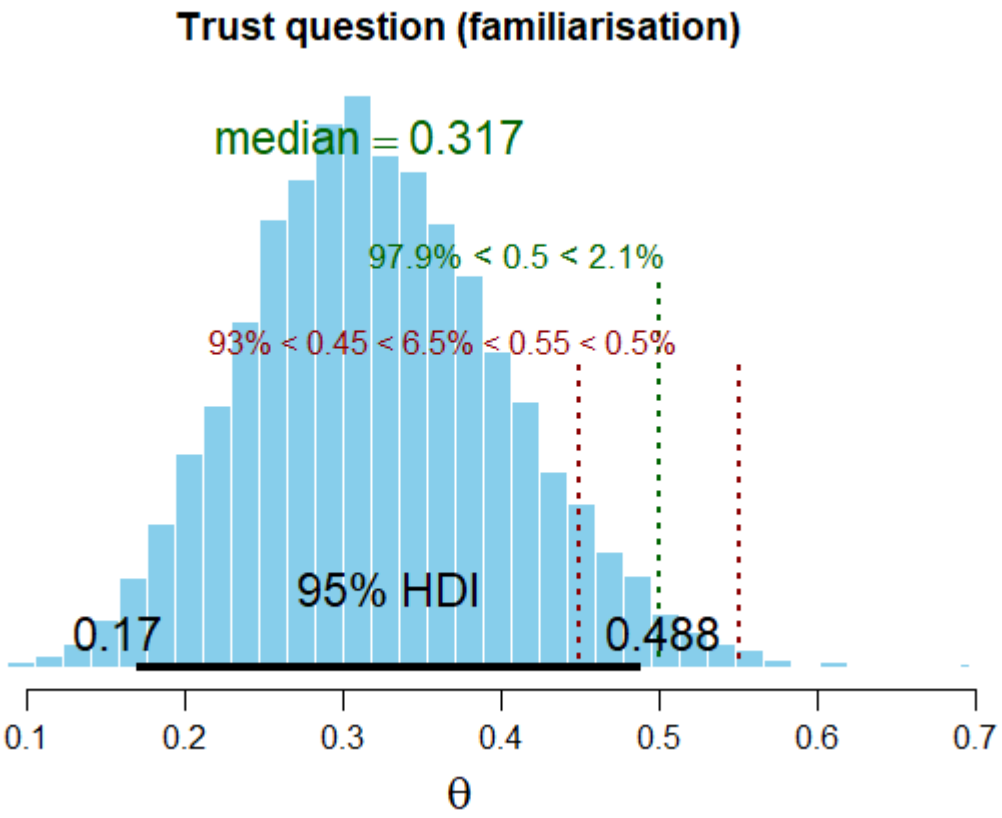

Figure 6.4. Posterior distribution showing no evidence for either hypothesis, i.e., it is not clear whether children selectively trusted an accurate informant, here coded as 1 (or an inaccurate informant; here coded as 0) more often than chance (0.50). The computed Bayes Factor for the alternative hypothesis is 1.79.

Test phase results

We tested the hypotheses that (i) children believe more often an informant whose claim was supported by a direct evidence (first-hand source: see) than an informants whose claim

was supported by an indirect evidence (second-hand source: tell); (i) once both informants are revealed to be inaccurate, they hold accountable more often an inaccurate first-hand source than an inaccurate second-hand source; and (iii) children selectively trust more an inaccurate second-hand source than an inaccurate first-hand source. None of the predictions were confirmed. As can be seen in Figure 6.5, the same proportion of children believed and hold sources accountable: nearly the same amount of children chose to believe (and hold accountable) the first-hand source (*N* = 13) and the second-hand source (*N* = 16); and nearly the same amount of children selectively trusted the inaccurate first-hand source (*N* = 13) and the inaccurate second-hand source (*N* = 16).

Figure 6.5. Amount of children who chose to believe, hold accountable and selectively trust the first-hand source (see) or the second-hand source (tell).

Since the proportion of responses were the same for the Belief and the Accountability question, we carried out a two-sided Bayesian Binomial tests based on the alternative hypothesis that the proportion of children who believed a first-hand source (as well as the proportional of children who hold a first-hand source accountable) was different from chance level (0.50).

As we can see from the posterior distributions plotted in Figure 6.6, the estimated median is 0.45, and the 95 % of the high density interval (HDI) [0.28 – 0.62] partially overlaps with the ROPE: more than 1/3 of the 95% HDI is contained in the ROPE (36.5 %). This suggests that we have inconclusive evidence about whether the data are better explained by the null or the alternative hypothesis. The evidence is inconclusive, as confirmed by the Bayes Factor (0.26).

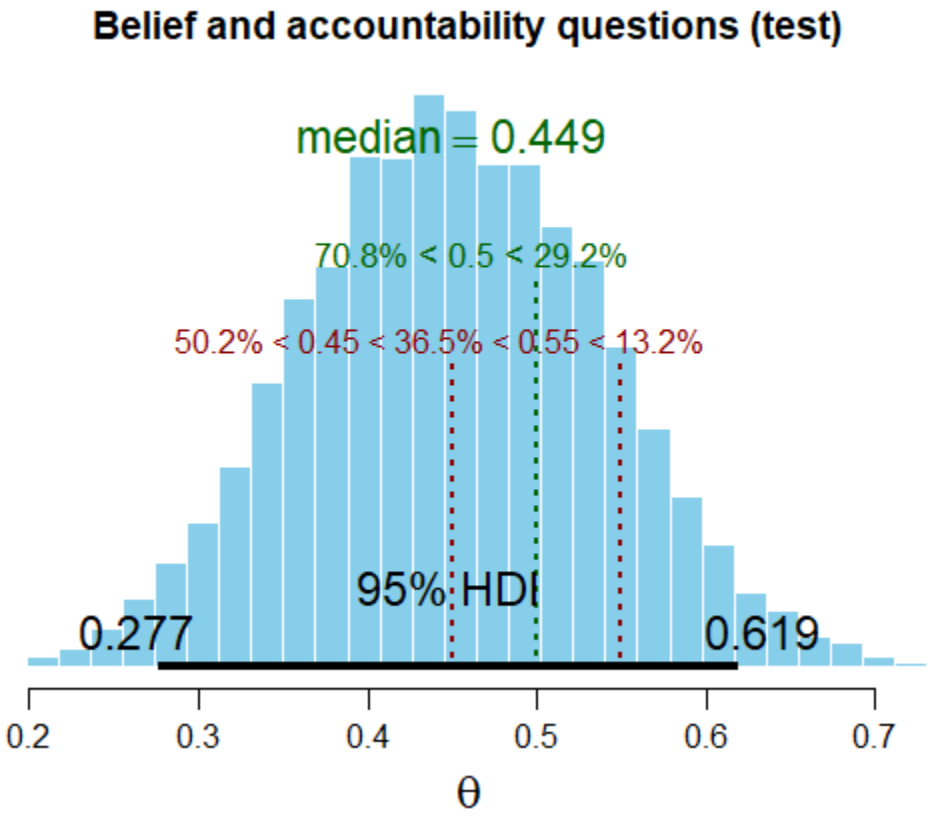

Figure 6.6. Posterior distribution showing inconclusive evidence for either hypothesis, i.e., it is not clear whether children believed (and hold accountable) a first-hand source, here coded as 1 (or a second-hand source, here coded as 0). The computed Bayes Factor for the alternative hypothesis is 0.26.

We finally carried out a two-sided Bayesian Binomial test based on the alternative hypothesis that the proportion of children who selectively trusted an inaccurate first-hand source was different from chance level (0.50).

As we can see from the posterior distributions plotted in Figure 6.7, the estimated median is 0.55, and the 95 % of the high density interval (HDI) [0.371 – 0.71] partially overlaps with the ROPE: more than 1/3 of the 95% HDI is contained in the ROPE (36.4 %). This suggests that we have inconclusive evidence about whether the data are better explained by the null or the alternative hypothesis. The evidence is inconclusive, as confirmed by the Bayes Factor (0.26).

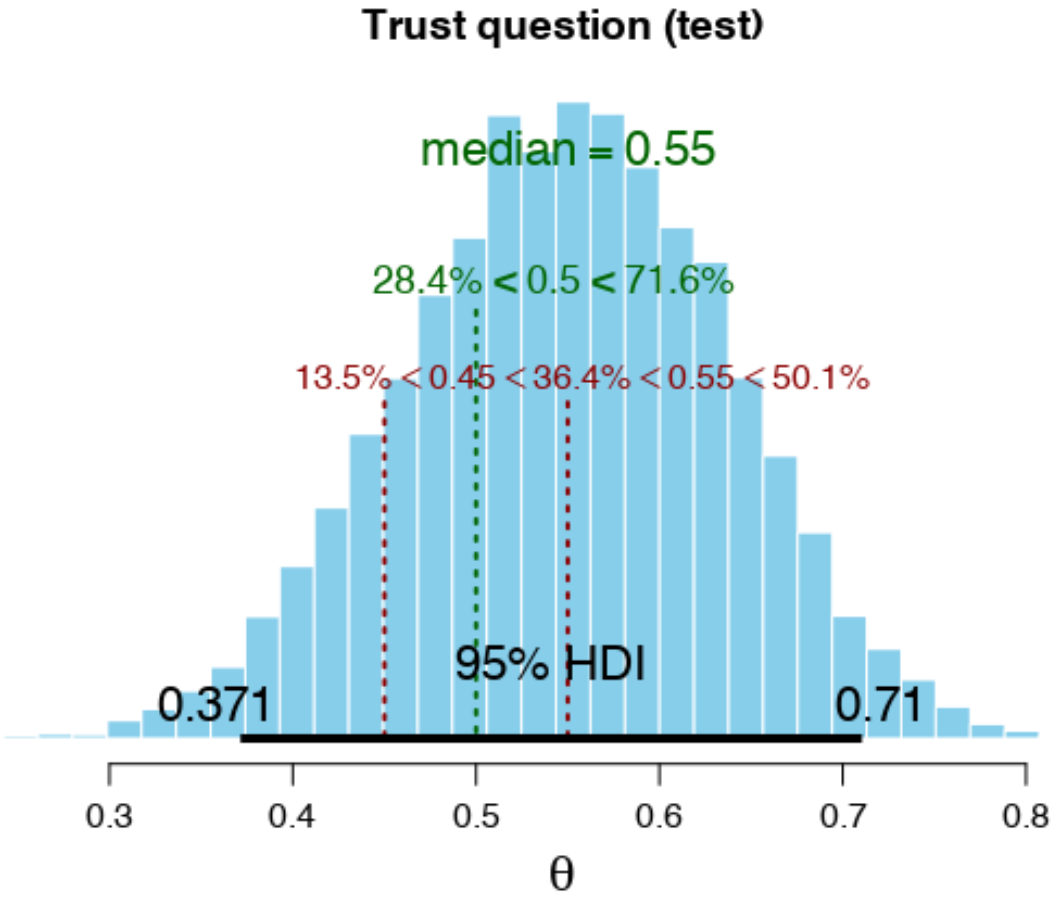

Figure 6.7. Posterior distribution showing inconclusive evidence for either hypothesis, i.e., it is not clear whether children selectively trusted a first-hand source, here coded as 1 (or a second-hand source, here coded as 0) more often than chance (0.50). The computed Bayes Factor for the alternative hypothesis is 0.26.

Discussion

Discriminating between reliable and unreliable communicators is an important skill to acquire to be able to retain useful and relevant information and discard useless and misleading one. One way to do so is to track communicators' commitment to the information transmitted. While we know that adults are sensitive to communicators' commitments, it is yet not known the extent to which children are able to track such commitments, and act upon them not only epistemically (by believing or not the information transmitted and its speaker) but also socially (by holding their speaker accountable). To the best of our knowledge, this is the first study investigating whether children hold inaccurate communicators socially accountable based on the type of evidential that the informants provided.

Our findings suggest that, when prompted to reward communicators, 6-to-8 years olds take into account their past accuracy in providing information; however, contrary to our

predictions, when prompted to hold inaccurate communicators accountable for misleading, children do not take into account the type of source (first- or second-hand) that communicators provided. In contrast with some other findings (Danovitch & Lane, 2020; Fitneva, 2008; Lane et al., 2018; Ozturk & Papafragou, 2016), children were also found not to discriminate between first-hand and second-hand sources also when asked what informants to believe (see Table 6.2).

Table 6.2. Summary of the results of the analyses conducted on children's responses to the questions in both the familiarisation and the test phase.

	Familiarisation Phase		**Test Phase**		
	Reward Question	*Trust Question*	*Belief Question*	*Accountability Question*	*Trust Question*
Distribution	24 accurate 5 inaccurate	9 accurate 20 inaccurate	13 see 16 tell	13 see 16 tell	16 see 13 tell
Median	0.84	0.32	0.45	0.45	0.55
95 % HDI	0.701-0.951	0.17-0.488	0.28-0.62	0.28-0.62	0.371-0.71
95 % HDI inside ROPE	0 %	6.5 %	36.5 %	36.5 %	36.4 %
BF_{10}	150.694	1.787	0.264	0.264	0.264

These results go against our prediction, but they can be accommodated with post-hoc pragmatic considerations regarding the paradigm. First, the dialogic interaction between the two informants may suggest that both informants are similarly confident about the information they are providing—or at least, confident enough to engage in an insisted discussion with the other informant, who expressed the opposite suggestion. Confidence in communicating a message was found to modulate commitment attribution ((Vullioud et al., 2017), and in our paradigm it may well have had a confounding impact. Second, children's reliance on the outcome of the game (i.e., the animal finding their food) is not particularly strong; this lack of reliance is one of the results of using an online protocol, which allowed for a third person rather than a first-person game. Without themselves relying on the information, children may be less motivated to evaluate of the evidence provided by the informants, while still be very engaged in the outcome of the search. However, the evaluations of the reliability of an agent may be beneficial for partner choice considerations, even if the reliability is inferred from a third-party observation. Finally, it could be that children interpreted the whole situation to be a guessing game rather than a social interaction between three animated characters. If that were true, the

presence of the desired object would have nothing to do with the testimonies of the informants, but it may rather be cued with the colour or the location of the boxes[24]; in this case, then, the testimonies should not be the primary evidence to consider.

This last explanation is corroborated by the results of the trust measure in the familiarisation phase. Contrary to any theory and previous findings in the literature (Koenig et al., 2004; Koenig & Harris, 2005), our results suggest that children tend to selectively learn more often from inaccurate than from accurate informants. We can speculate that instead of looking for the desired object, children prioritise collecting additional relevant information regarding the inaccurate informants—specifically, whether they manifest a stable trait of incompetence. In fact, avoiding incompetent or malevolent partners and communicators may be more important that choosing competent and benevolent ones. However, it is easier to explain this pattern to responses without referring to partner-choice strategies, and rather by using of a simple heuristic: if the one informant was accurate before, then the other informant must be accurate now. If children engaged in this type of reasoning, it seems very likely that they interpreted the situation as a guessing game and not as a socio-communicative interaction.

One possible criticism of the current study is that children's responses may not depend on how much communicators are perceived to modulate their commitment to what they communicate, but rather on their evaluation of different ways of acquiring knowledge. If first-hand evidence is evaluated epistemically more valid than second-hand evidence, then communicators will not be perceived as more or less committed to the information they are providing, but merely committed to different type of information (to have seen, and to have been told). However, if the desired object is not found in the location mentioned by the informants, the one informant who deferred their knowledge to another (mistaken) agent should not be held accountable for the misleading information. Therefore, while this argument may very well be true, it does not provide a better explanation for our pattern of results.

We cannot of course exclude the possibility that our finding is, on the other hand, diagnostic of a lack of capacity of tracking such commitments in children. However, given the substance of the concerns raised regarding our online paradigm, a different protocol in which the dialogues between agents is minimised, and in which children's reliance on the accuracy of the information provided is higher may help settling the issue.

[24] Some of the comments from the children who participated in the study suggest that these dimensions were privileged when they were prompted to make choices.

Part IV. Practical implications

As discussed in the previous chapters, the notion of commitment has been used by researchers from different fields, such as philosophy of language, social ontology, game theoretical approaches, moral philosophy. My approach regarding the psychological underpinning of commitment shows that partner's reliance is the main evidence to consider when investigating the perception of commitment.

In the introduction I spelled out this hypothesis, by arguing that most social interactions require people to be able to recognise the situations when it is beneficial to rely on one's behaviour (or on each other's contribution to a joint goal). Perceiving others to be committed to something helps to recognise such situations and sooth the uncertainty related to others' actions. People perceive commitments to be in place when there is evidence that one is going to do X that the partner can rely on, or that the partner is going to rely on one's doing X—in other words, of a *commitment-reliance relationship*. Critically, such evidence can be minimal: even minimal cues can be taken as evidence that a commitment is in place, and reduce the uncertainty about others' future behaviours.

Chapter 1 and Chapter 2 presented two sets of studies investigating these minimal evidence, and specifically how implicit cues of a commitment-reliance relationship trigger normative and non-normative expectations about agents' behaviour. In Chapter 1 I showed that the costs paid in order to take part into a joint activity enhance commitment attribution, and so does a history of successful interaction (Bonalumi et al., 2019). In Chapter 2 I showed that the knowledge that a partner is relying on you doing something is enough to trigger a perception of a commitment being in place, even in the absence of any 'confirmation' or ostensive signal that one intended to do something (Bonalumi et al., 2021).

Collectively these findings support my general hypothesis that people perceive agents to be committed when others are relying on them. Critically, we found that participants' expectations about others' behaviour were not merely predictions or "hopecasts", but also normative expectations: participants judged agents who did not live up to partner's reliance as more blameworthy, less trustworthy, and more obliged to provide apologies or explanations. Contrary to what was found by Michael and colleagues (2016b) in a vignette study on the effect of coordination on the sense of commitment, we found that affective and normative measures strongly correlated: the sense of commitment triggered by minimal cues of reliance already entailed a normative stance.

The fact that partner's invested cost and repeated interactions have (perceived) normative consequences, but coordination had not, is consistent with other findings in the literature. John Michael's research with robots shows that coordination is a milder cue of a commitment-reliance relationship compared to invested costs and time. In fact, a humanoid robot's effort investment in a joint action motivated participants to live up to their (robotic) expectations (Székely et al., 2019), whereas coordinating together with a humanoid robot did not elicit the same motivation in participants (Vignolo et al., 2019). Furthermore, when mutual knowledge about coordination was experimentally manipulated, coordination lost its magic (McEllin et al., 2022). Thus, coordination may not enhance a sense of commitment because of the type of co-representations that it elicits, but rather because it is perceived as a kind of effort, however mild. I predict, I believe consistently with Michael, that increasing the difficulty and complexity of coordinating would cue a stronger commitment-reliance relationship; this would, in turn, increase its effect of reducing uncertainty about agents' behaviour, and thus enhance both commitment motivation and the perception of commitment.

Furthermore, our findings show that minimal cues of commitment-reliance lead to similar patterns in different types of social interactions: Chapter 1 depicted cases in which an agent failed to accomplish a joint goal, while Chapter 2 depicted cases in which only one agent was expected to follow through a commitment that their partner relied on. In the cases depicted in Chapter 1, the interests of the two agents are symmetrical, thus the roles of 'the one who commits' and 'the one who relies' are confounded: both agents share the same goal, and have similar roles in achieving of the joint goal[25]. This is a frequent case in our social interaction, and such cases have been the focus of social ontologists (Bratman, 1987; Gilbert, 2014; Searle, 2010; Tuomela, 2007), John Michael's research on the sense of commitment as well as Mike Tomasello's research on (the developmental basis of) collaboration and joint commitments (Michael, 2022; Tomasello, 2009). In the cases depicted in Chapter 2, instead, agents are involved in an asymmetrical dynamic, in which the costs of defection will be particularly harsh for only one of the two parties. Our results point out that the factors cueing commitment-reliance in such cases have the same effects in situations in which interests are or are not symmetrical, and 'the one who commits' has the same or a different role than 'the one who relies' in the commitment-reliance dynamic. This suggest that in cases of asymmetrical

[25] Of course, sharing a goal does not entail that the agents must also have similar causal roles in achieving that goal (Gilbert, 2009; Searle, 1990; Tuomela & Miller, 1988); this was however the case for the scenarios described in the experiments from Chapter 1.

commitments, the kind of cognitive processes at play are the same as when agents are involved in joint commitments.

Commitments help solving this temptation problem, i.e., the problem of relying on others doing something despite the potential opportunities for others to defect. Also in communicative context, commitments help solving the same problem: more specifically, the problem of relying on what others had communicated, despite the potential opportunities for them to deceive. I claimed that people would perceive others to be committed to do something (beneficial for them, or a joint goal) in view of the evidence of commitment-reliance relationship that is provided by the agents or that is salient in the common ground. Such evidence can be evidence of one's willingness to keep a promise, but it can also be evidence of one's communicative intentions to reassure a partner about a state of the world.

Relying on communicated information, i.e., changing your course of action on the basis of the information received, should affect the perception of commitment in a similar fashion than it does with respect to future actions. Chapter 3 and Chapter 4 explored this question. In Chapter 3 I presented a set of studies showing that participants would modulate the perception of commitment on the basis of a partner's reliance and not on the basis of the explicitness of the speaker's promise (Bonalumi et al., 2020). In Chapter 4 I presented a study showing that partner's reliance also affects plausible deniability (Bonalumi et al., 2022). Both sets of studies showed that the perception of commitment is not influenced by the level of meaning (i.e., whether the misleading content was explicit, enriched, or implicated). However, the level of meaning impacted the perception of promise being violated (Chapter 3) and plausible deniability (Chapter 4).

These empirical results contrast with theories according to which commitment to what is explicitly asserted is higher than commitment to what is implicated (Morency et al., 2008; Reboul, 2017) by highlighting that at least one major factor (partner's reliance) can overshadow the boundaries between implicit and explicit. Even more in contrast with some literature that credits processing differences among different types of implicit contents, i.e., different levels of meaning (Doran et al., 2012; Franke et al., 2020; Noveck & Posada, 2003), we found that the level of meaning did not impact speaker's accountability in either of the two sets of studies presented in Chapter 3 and Chapter 4—while reliance consistently did (Bonalumi et al., 2020, 2022).

Reliance had a robust effect on how people hold others accountable. We found it to affect also the scope of speakers' strategic uses of language: denying having meant something when

the audience relied on something occurring was judged to be less plausible compared to cases when the audience was not relying on (Bonalumi et al., 2022). This finding supports cognitive approaches that describe denials as attempts for context reconstructions (Mazzarella, 2021); it also enriches the experimental literature on plausible deniability by showing that not all denials are equally effective, and that at least one factor (partner's reliance) systematically affects deniability (Lee & Pinker, 2010).

While the level of meaning did not have any effect on speakers' accountability, it did influence plausible deniability: denying having meant an implicature was deemed more plausible than denying having meant an enrichment. This difference resonates with previous theoretical and empirical contributions (e.g., Brown & Levinson, 1987; Doran et al., 2012) and hints to the idea that deniability might be linked to a naïve notion that language is a digital medium—discrete, context-independent, and carrying meanings that are fixed (Pinker, 2004).

Theoretical and empirical work on the naïve notion of lying is vividly debating the issue of whether implicated content is taken into account when judging whether people lied (Antomo et al., 2018; Borg, 2017; Meibauer, 2018; Viebahn, 2017, 2019; Weissman & Terkourafi, 2019; Wiegmann et al., 2016, 2021; Willemsen & Wiegmann, 2017). More recently, it has been proposed that a definition of 'lie' should be commitment-based (Marsili, 2021b; Reins & Wiegmann, 2021). While our findings cannot decisively inform such debates, they are relevant to the extent that promises are reducible to assertions and thus a broken promise may be reducible to a lie (Marsili, 2016). In fact, I showed that the level of meaning had no effect on perceived commitment, but it had nonetheless an effect on perceived promise violation: promises conveyed via an enrichment were judged to be 'broken' (Study 3a; pp. 80-87) while promises conveyed via an implicature were judged otherwise (Study 3b-3d, pp. 87-98). If breaking a promise and lying are viewed as similar, then our findings suggest that the level of meaning is impacting the folk notion of lie.

Contrary to this interpretation that promises and lies are perceived as similar phenomena, Yuan and Lyu found that implicit promises and implicit assertions brough about different effects in terms of speaker's accountability (Yuan & Lyu, 2022). Replicating our findings in Mandarin Chinese, participants judged speakers who had broken promises to be equally blameworthy irrespective of whether the broken promise was implicit or explicit; and, similar to Mazzarella and colleagues' findings (2018), speakers were on the contrary judged less severely when they conveyed false implicatures compared to when they assert something false. Such interaction between accountability and speech act is intriguing, but it is unclear how

much of this finding is due to a manipulation of partner's reliance (as the authors admit; see Yuan & Lyu, 2022, p. 139).

The contrasts in the literature show that what is evaluated as 'explicit evidence' may differ depending on the circumstances. While typically a verbal assertion is considered 'explicit', there might be circumstances (e.g., when signing a contract for buying a house) when such explicitness is not enough for the parties to be reliant on it. In any case, the potential impact of our results on debates on plausible deniability and folk notions of lies and promises proves that commitment and reliance are the neglected factors that researchers interested in communication should focus on (see Geurts, 2019).

That communication and joint action (may) have normative consequences is recognised by children as young as three (Rakoczy et al., 2008; Schmidt et al., 2016; Tomasello, 2018). While children at that age were already found to pay costs in order to allow their partner to obtain their share of the joint goal (Hamann et al., 2012), they tried to re-engaged with a defective partner more often when an agreement was in place (Gräfenhain et al., 2009). Also, three-year-olds complained more when a partner was intentionally interrupting a joint activity (a game) because lured by a tempting alternative compared to when they did so because they did not know how to play the game or because the apparatus of the game broke (Kachel et al., 2018). Furthermore, children at three were also found to excuse defective partners more often when the partner acknowledged that they were leaving, or asked permission to leave (Kachel et al., 2019).

Such results persuaded us that children would be able to discriminate between situations in which a partner abandoned a joint activity because of a selfish motive (i.e., to play another tempting game), and because of a moral motive (i.e., helping someone in distress). We predicted that children would release partner more often in morally motivated cases: we were wrong. Children did not show different reactions in such cases, which of course is not a proof *per se* of a lack of such capacity; but it shed some doubts about children's ability to evaluate the type of justifications they are provided (or they can provide). Consistently with this idea, Patricia Kanngiesser (2021) found that children with 3 years olds (but not 5 years old) judge it to be more permissible for an agent to break a promise for selfish than for moral motives.

Commitment violation can be related to a failed contribution towards a joint goal or a missed future action, but also related to the transmission of a false information. A relied on information that is later found to be false impacts children's evaluations of its source: inaccurate sources are less believed and less rewarded that accurate ones (Koenig et al., 2004;

Koenig & Harris, 2005; Ronfard et al., 2019). A more fine-grained appreciation of the stakes involved when speakers convey information, and more specifically when they are expected to convey relevant information, include evaluating the source of one's claim. Thus, people should not only to believe, but also blame more speakers who suggested something false by referring to a first-hand claim than a second-hand claim. The evidence provided by the communicator is much stronger in the first than in the latter case: to 'have seen' X raised an array of expectations—that the communicator is certain about something; that the communicator was present while X, and as such has an epistemic authority over the knowledge of X; that the communicator is willing to put their reputation at stake for the truth of X. Adults' reactions to such commitment violations were consistent with our hypothesis: participants believed more, and blamed more communicators who stated 'I saw that X' than those who stated 'I was told that X', in the event that X is false (Mahr & Csibra, 2021).

The findings in the literature about whether children discriminate between these source claims are not conclusive (Aboody et al., 2022; Fitneva, 2008; Koenig, 2012; Ozturk & Papafragou, 2016). Our results line up with these findings by suggesting that investigating this capacity in children mobilise a lot of pragmatic inferences that can override a cognitive phenomenon already present. Furthermore, our online paradigm was prompting third-party evaluations, and it could be that third-party judgements (in particular, children's third-party judgements) are different from second-party judgements—such as those following a commitment violation when children themselves are the ones paying the costs for. In fact, I would predict that the higher children's reliance on the information provided, the more tuned their appreciation for different source claim would be. Further research will shed lights on this phenomenon.

Finally, the final two sections present some speculative arguments. I will discuss how the results of the studies presented in the previous chapter have implications not only for the current academic debate, but also in the ethical and legal domain; more specifically, how our theoretical and empirical contributions can inform the current debates about sexual consent and about digital communication and misinformation.

Some implications about sexual consent

Consent is one foundational principle in democratic societies and a primary tenet of how individual liberty and dignity are warranted and protected by the law. Contracts vicious of consent are nullified. Services requiring people to be subjected to a third-party treatment or action (e.g., medical treatment, participation in an experiment, tracking of the cookies in our browser), are conditional on consent—more specifically, 'informed' and uncoerced. Sexual consent, i.e., the consent in engaging in a specific sexual activity, is a particularly thorny issue. In many jurisdictions[26], the lack of sexual consent becomes the ground for rape charges. What is considered a lack of consent, is however, not trivial. Much of the academic work on sexual consent shows that there is no widespread agreement on the definition of consent (Beres, 2007).[27] This lack of consensus brings problems for policy makers, legal theorists, and activists to properly fight sexual violence.

A fruitful approach to better understanding (some) sexual violence comes along with a better understanding of how consent is both given and interpreted: it is of key importance to unveil what are the psychological attitudes that people manifest with regard to consent, so to assess whether the normative and legal principles can be easily taken in—as coherent with our interpretative processes and how we perceive others to consent to something[28]. My contribution to explaining reliance and commitment cannot address the complex issues of

[26] Consent-based norms, i.e., laws for which rape charges require a lack of consent, are mostly adopted by Commonwealth countries, some US states, and few other countries (including Germany and Turkey). Most of European legislations adopt a coercion-based norm, i.e., laws for which rape charges require a proven presence of coercion, violence, or threat. Strikingly, some countries adopt a consent-based norm but do not criminalise intramarital rape, unless violent.

[27] Consent is at times viewed both as a behaviour and a mental disposition; as the criterium discriminating between good or morally acceptable sex and bad or morally unacceptable sex; as any agreement rather than free agreement only.

[28] Roseanna Sommers (2019) makes a similar point. In her studies, Sommers shows that the normative notions of consent (Beres, 2007; Dripps, 2009) and the naïve notion of consent are not fully overlapping, with the latter being largely more liberal than the former. Participants reported to understand situations in which consent was extorted with deceptive means to be still situations 'consented to', i.e., consent obtained via deception is still considered consent—unless the level of deception in coercing a consent is such that it *transforms* the consented activity in something intrinsically different. For example, deceiving a partner by falsely stating that one is unmarried does not make the subsequent sexual interaction unconsented, even if the partner would have not given consent if they knew that one is actually married; deceiving a partner by falsely stating that one is their twin brother, instead, is perceived to transform the subsequent sexual interaction into something different, and thus not consented to (Sommers, 2019, pp. 2289–2290). Of course, it must be remembered that folk intuitions (as well as many salient contextual assumptions) may themselves be the results of the very same power structures that activists and scholars are fighting.

systemic sexism and power dynamics at play, and definitely it cannot provide normative alternatives to the consent principles. But it can shed lights on what processes may be in play when people engage in interactions that partners did not consent to—but may have been *perceived to have implicitly consented to*.

That consent is taken to be the criterium for legal treatment suggests that consenting bears normative implications: to consent to X is to agree that X will occur, accept the related consequences of X but also accept the obligations that arise with the fact that X has to occur. I claim that consent (explicit or implicit) is *perceived* to have such normative power. Consent is about committing: to consent to X means to commit to X. Consent is about relying too: to consent to X means to rely on the fact that other recognises that the scope of the consent is not unrestricted.

Our findings (Bonalumi et al., 2019, 2021; see Chapters 1-2) show that people perceive others to be committed to do something, even in the absence of any verbal cue that establish their intention to do something. The mere belief that a partner relies on an individual to comply to their expectations is a strong enough cue for participants to ascribe commitment to that individual. The mutual knowledge of the partner's effort, the repetition of a routine, and even one's silence were taken to be evidence that one was in fact committed to comply to the partner's expectations. These results show that people understand others to *implicitly* consent to engage in an activity—and that when consent to X is explicit, they may understand others to have consented also to a range of scenarios that are (or may be) implied with X. But our results suggest also that people would also hold others accountable for raising such expectations, and if consent is withdrawn, they may perceive it as a commitment violation and thus entailing social costs. The practical implications of these findings are extremely problematic and cannot afford to be ignored.

What in theories of commitment and communication is described as "how people interact" have thus tremendous consequences in real life. People may misinterpret the scope of an implicit consent, or even its presence. And most likely this misinterpretation is translated in an overestimation of such scope, which gives people arguments for excusing marital rape, rape during sleep, and the imposition of unwanted sexual acts. Pineau's communicative model of sexuality (1989) and the explicit consent principle attempts to put a boundary to this misinterpretation. Kitzinger and Frith (1999) and O'Byrne and colleagues (2006) present arguments supporting the idea that people are competent in recognising cues of refusals (an absence of consent), but it is unclear whether this competence would apply as easily to the

scope of a consent. In fact, we know that communication does not typically follow the path of making things explicit.

According to Pineau's influential proposal, a consent is given when there are explicit cues of one's willingness to engage in sexual interactions. By giving and requesting an explicit consent, any ambiguity about a consent being present is (should be) removed. The benefits of implementing such habits are evident, and they outweigh the costs: instances of rape due to unconsented interactions would be disincentivised, and accountability in such cases would be better tracked and not negotiable (either a sexual interaction was consented to, or it was not). Such principle is likely to be the best principle to ground legislation on, because it is protective of vulnerable groups—those who would find themselves more often in the situation of being falsely ascribed a consent. However, it, comes with some prices[29]. While a culture of consent is being (luckily) promoted in education and legislations, it may still fall short regarding how people interact in sex (at least, in heterosex; see e.g., Beres et al., 2004). Such considerations are enriched by the fact that some may experience sexual interaction and desire not as something to plan but something to discover (Angel, 2022); entailing playing and innuendos that contribute to build up the feelings of safety that sexual interactions should be grounded on[30]. Furthermore, and even more profoundly, it may be easier for the audience to implicitly signal their (un)willingness rather than confronting explicitly a sexual request. Such ease in cueing consent and unconsent in grades of explicitness depends (also) on how we communicate in general.

When we communicate, we rarely make everything fully explicitly; we rather make things explicit enough, in an optimal way, only to the extent that we think is necessary for the audience to understand what is meant. Communication is not about decoding a series of auditory stimuli, but is about making inferences about what the speaker intends us to know

[29] One shortcoming is that even an explicit consent can be an unwanted consent. When hierarchies are present or perceived, the willingness to comply is too easily triggered by the explicit request to give consent to unwanted situations (see Sommers & Bohns, 2018, for examples beyond sexual consent). Some power relationships are so unbalanced that they do not give a comfortable space for an explicit consent not to happen. Furthermore, the consent principle assumes for instance that people are aware in advance of the type of interaction they might want to build; that consent must be then explicitly and promptly expressed or retracted whenever a sexual interaction is negotiated; but more problematically, it can reinforce the (gendered) idea that sex is a kind of interaction in which one part is proactive and the other part is consenting (Angel, 2022). While these points worth debating, they go beyond the scope of this chapter.

[30] Of course, the way in which people do interact in sex, as well as the idea that flirtations, innuendos, and foreplay can be initiated *only* when ambiguity is present are themselves influenced by centuries of patriarchal habits; but the idea that they *must* be initiated explicitly can crash with the process of feeling safe in engaging in a sexual interaction.

(Sperber, 1994; Sperber & Wilson, 1986/1995). Such inferences are produced when communicated stimuli take the form of verbal and explicit utterances as well as implicatures; what the audience is aiming for is, anyway, the speaker meaning.

In a scenario in which two individuals are negotiating whether they want to be involved together in a sexual activity, an audience can infer a speaker meaning (or communicator meaning) such as "I consent to the activity X" based on multiple types of evidence: verbal evidence (the communicator asserts "yes, I consent to X"); and/or other kind of non-verbal evidence. Non-verbal evidence includes mutually manifest cues of consent that can be more or less explicit—ranging from ostensive signals of agreement to the shared history of an interaction. In many cases, non-verbal evidence is interpreted to be explicit enough so that giving or asking for an explicit consent could feel redundant.

In situations in which a pleasant interaction builds up, it is (or feels) often obvious and manifest that the parties are agreeing on what is happening. This suggests that making things explicit in some situations may be perceived redundant. Furthermore, making things explicit can be perceived as a lack of trust (Chennells & Michael, 2021), and may require a re-negotiation of the relationship itself (Thomas et al., 2014, 2017).

Consent theorists agree that some cues to consent are explicit though not verbal. Nodding, initiating the removal of one own's clothes, and other non-verbal consents can be considered as genuine consents (Pineau, 1989). However, some cues that we found to impact the perception of commitment are very much problematic. Specifically, we found that a history of doing something together lead participants to expect agents to be committed to do that something in the future. Such expectation is not a mere prediction but is also a normative attitude: agents who fail to comply to such behaviour are perceived to have misbehaved.

The implication of these attitudes in the context of sexual consent are evident: people may have an intuitive understanding of contextual assumptions such as 'we used to do X in context Y' that may lead them to perceive their partner committed to do X, even in the absence of any other cue of their partner's willingness to do so. In view of the evidence that an interaction was repeatedly consented over time, people may hold partners committed to that interaction in the future, as if they provided a blanket consent for similar interactions in the future.

A blanket consent is conceived as a broader consent for a less specified range of future actions and practices. Such idea assumes that the scope of what you are consenting to is broader than the one initiated by the consent itself. Agreeing to be a wife, or a partner, or merely having been repeatedly involved in a sexual relationship can be perceived as a cue for a blanket consent to meet certain expectations and be committed to similar activities as

previously occurred. Such situations (i.e., rapes[31]) are (frightenedly) much frequent. A majority of women reported to have experienced sex (rape) while asleep[32], and the notion of marital rape is still treated as an impossibility in some legislations (e.g., India and Egypt among others), and was considered as such in most legislation until recently, as rape was considered first a crime against a man's property (a father, or a husband were victims of a theft), and then a crime against a woman's honour and reputation and not against the person. Being involved in a marriage is perceived to be evidence for certain obligations—in marriage such obligations are formalised as conjugal rights.

Both rape-during-sleep and marital rape are instances of a misinterpretation of the scope of consent, and a forceful application of a blanket consent to activities that one has not consented to. However, in many other circumstances the idea of a blanket consent is less dangerous and seems very intuitive. For instance, if two individuals agree on having sex, they will most likely be perceived to agree to kiss too. Some contextual assumptions, including the ones that are patriarchal heritages, will arise as salient for individuals to interpret specific factors as commitment-reliance cues.

When cueing a commitment-reliance relationship, such as one's willingness to engage in a sexual interaction, one is not committing to such engagement. But they may be taken to be committing to it. Expressing explicit consents in such circumstances may be useful to disambiguate the scope of the consent itself, and disincentivise (some) sexual violence to occur. However, resizing the scope of consent could also occur by removing salience to some contextual assumptions that are (unfortunately) still too present in our societies.

[31] I believe that imposed sex based on the history of past interaction is rape; however, I am not arguing that all rapes are caused by overestimations of the scope of the consent.

[32] ***https://www.theguardian.com/society/2021/jun/15/the-sexual-assault-of-sleeping-women-the-hidden-horrifying-crisis-in-britains-bedrooms*** (The Guardian, 2021).

Some implications about digital communication

Online interaction on social media is currently one of the most widespread means of communication (Newman et al., 2018). Compared to mere online communication, social media allow people to interact also in a new and different fashion: by posting contents online, reacting to those contents, and reposting those contents. These new ways of communicating[33], such as online posting, microblogging, and retweeting are already having a huge impact in society—shaping the way in which private and professional relationships are now handled, and enabling the spread of misinformation in ways that are more visible than ever (Altay et al., 2021; Guess et al., 2019). While there are some similarities with the way in which 'real' communication work, the differences lead to some challenges for present accounts of communication.

It is of primary importance to understand not only what kind of communicative acts they are but also what responsibility we hold users to maintain and what modulates such responsibility. I will articulate how our theory of commitment predicts that the perception of commitment in online contexts can sometimes be even higher than in conversational contexts, using the example of two kinds of online behaviours, i.e., microblogging (posting or tweeting short texts, links or images), and retweeting (reposting or sharing other users' contents).

Microblogging is a new kind of communicative act that involves broadcasting original contents whom authorship is the user themselves. As such, microblogging appears to be a proper speech act, whose illocutionary force differs across situations: similar to how people speak *tête à tête*, by posting users can assert, ask question, apologise, and so on. It is not clear, however, who are these communicative acts for, namely, who is the audience of such posts. The unrestricted scope of the audience makes microblogging look more like production of content such as press (intellectual) or artistic content. Retweeting, or sharing other users' contents, is even trickier: it does not seem to have a counterpart in *tête à tête* communication, and while it may be assumed that shared contents are *endorsed* or *approved* contents, it is not a trivial interpretation—as Marsili (2021) notes, many followers retweeted the legendary 'covfefe' tweet from US President Trump[34], but none of these were interpreted to be endorsements of the conveyed content. Retweeting shares some relevant aspects with quoting (they both reproduce another content) and with indicating (they both direct the audience's

[33] There is already important work about whether these new communication means are substantially new kind of communicative acts; see e.g., Carr and colleagues (2012), Marsili (2021a), as well as Xie and colleagues (2021).

[34] ***https://web.archive.org/web/20170531054122/https://twitter.com/realDonaldTrump/status/869766994899468288*** (Trump, 2017).

attention to a content), but also differs in some important details. Retweeting reproduces a content that is identical to the original one, but also duplicates it and redirects the duplicated content to a new and different audience, offering them the possibility to interact with such new content (Adamic et al., 2016).

What are the social and epistemic responsibilities of digital users? Are these responsibilities the same as the ones borne by offline communicators? Are digital users perceived committed to what they post or re-tweet as they would be offline?

On the one hand, lots of phenomena happening online are uncontrolled. For instance, doxing[35], vote brigading[36], and cyberbullying[37] are typical cases of uncontrolled behaviour for which the users are not held accountable. In particular, but not exclusively, when anonymity is warranted, users can get away with hostile behaviour and spreading outrageous contents in a way that is not possible during face-to-face interactions. The theory of the strategic speaker (Lee & Pinker, 2010) predicts that the 'interpretative wiggle room' can and are very easily applied to such situations: exactly because often audiences are unrestricted, the negotiation of what was meant is easier, and denying to have meant something problematic, false or offensive, comes as costless (see also Dinges & Zakkou, 2021; and Guercio & Caso, 2021 for similar approaches to dogwhistling). In describing the different denial strategies available to speakers, Boogaart and colleagues (2020) presents several examples of denials from online communication: these denials span from denying having meant a certain implicature to appealing to irony and other non-literal meanings. Denials that exploit irony are particularly efficient in online contexts, as they allow the speaker to avoid reputational consequences while still permitting their message getting through the target audience.

On the other hand, given the public context in which digital communication occurs, such denials should also be less efficient. And sometimes they are indeed less efficient: at times social media behaviour is monitored and punished. Audiences often assume that users agree and endorse what they retweet, rather than interpreting retweeting as a distancing cue; and they hold users committed to such endorsements or to the literal meaning of ironical assertions (often appropriately). There is a multitude of cases of people getting fired over posted or

[35] See for example the case of the Reddit vigilantes of the Boston Marathon bombing, ***https://web.archive.org/web/20131215111626/http://www.3news.co.nz/Innocents-accused-in-online-manhunt/tabid/412/articleID/295143/Default.aspx*** (2013 April 22, 3 News).

[36] ***https://www.wired.com/2016/09/imdb-voters-tanking-indies-theyre-even-released/*** (WIRED, 2014 Sept 14).

[37] See Santre (2022) for a review of the phenomenon.

retweeted unacceptable jokes[38], or being publicly asked to account for the implications of fake or unacceptable content or fake news[39], without mentioning the woke movement against digital bullyism.

Thus, theories of communication predict that the epistemic and social responsibilities should be very high, because of the public setting in which communication occurs—the more public the communicative act is, the higher the reputational consequences should be (Altay et al., 2020). On the other hand, the use of second-hand evidence, as well as the use of irony or other dogwhistlers should be taken as a cue of speaker distancing from the content (such as it happens e.g., in gossip). The balance between accountability and plausible deniability seems to be even fuzzier than with actual conversations.

What modulates this balance between accountability and plausible deniability in online contexts? Are digital users perceived as responsible for the cognitive effects that online content, either produced or shared, has on their audiences? Our theory of commitment in communication provides us with tools to (start to) answer this question, and to explain how and why ascriptions of responsibility can be particularly severe.

Digital behaviour is inherently ostensive (Marsili, 2021a). As any instance of ostensive communication, it carries a presumption of relevance (Sperber & Wilson, 1986/1995): namely, the audience will assume that the user had posted or shared content in such way that the content will yield cognitive benefits at a cost that is optimally geared (not higher than necessary). When exposed to communicated content, the audience will retrieve contextual assumptions that will help them restrict the array of interpretations that may arise and discard the ones that do not reach an optimal relevance (whose yielded cognitive benefits are lower than the invested cognitive effort). Like other types of communicative behaviour, digital behaviour raises expectations in the audience about the relevance of that content.

The relevance of what is communicated is largely influenced by whether and to what extent the audience relies on it. The studies presented in the present dissertation (in particular, the studies presented in Chapters 3-4; namely, Bonalumi et al., 2020, 2022) suggest that when

[38] See for example Justine Sacco's case, ***https://www.nytimes.com/2015/02/15/magazine/how-one-stupid-tweet-ruined-justine-saccos-life.html*** (The New York Times Magazine, 2015 Feb 12).
[39]***https://edition.cnn.com/2020/06/28/politics/trump-tweet-supporters-man-chants-white-power/index.html*** (CNN, 2020 June 29).

the audience rely on what is communicated (or what is understood to be communicated), the more the reliance, the more communicators are held accountable for the cognitive effects they cause on their audience. Even more important for our purposes, our studies show how the manipulation of contextual elements (which are part of the common ground) have an influence on communicator's accountability. Participants hold communicators differently responsible for the same communicative act in different contexts (because different contexts influence the expectation of relevance of such act). The negotiation of the common ground ("I didn't know that…"), in fact, plays a big role in plausible deniability and in mitigation of social costs.

While in dyadic interactions establishing common ground can be achieved (more) easily, when interactions are translated in a digital platform this is less obvious, since the audiences of a given digital content is not always the audience that a user had in mind when producing such content—the actual audience and the target audience are different.

The target audience of a digital content is the assumed audience for whom a user is performing a digital communicative act. For example, when researchers tweet about their new pre-print, they do it for other researchers who share the relevant common ground—the cultural and scientific environment in which such pre-print would be relevant, and the shared values that are embedded in it. However, the actual audience that end up being exposed to such tweet may not share the same values and the same cultural and scientific assumption. The example of Nicolas Baumard' Twitter advertisement of one of his work together with Lou Safra, Coralie Chevalier and Julie Grèzes (2020) is particularly emblematic. In fact, their advertised Nature article was furiously criticised over Twitter, leading the senior author to delete his own account after an avalanche of outraged rebuttals and insults[40]. The criticisms that the manuscript (and the authors) received were the result of (a) a communicator who posted a series of tweets meant to be for an academic audience; and (b) an actual audience who did not share the same common ground with the communicator (in this case, enough familiarity with e.g., research methods and academic agenda and jargon; or sometimes familiarity with different academic agendas and jargons). The contextual assumptions raised by the target audience of Baumard's tweets were very different from the ones raised by the actual audience of his tweets.

This mismatch raises some issues about what contextual assumptions the actual audience will retrieve in order to make sense of the communicated content as optimally relevant. The fact that there is a lack of common ground between user (communicator) and audience will easily lead to an interpretation of the communicated stimulus that may not be matching what

[40] See for example ***https://quillette.com/2020/10/03/time-and-perceptions-of-trustworthiness-the-row-over-a-novel-study/*** (Quillette, 2020 Oct 3).

the user meant: and thus, the implications of any digital behaviour may be various—as each audience may end up referring to different contextual elements that were not initially salient.

The risk that a digital content may raise expectations of relevance that lead to different interpretations depending on the different contextual assumptions that are salient for the different audiences bears important consequence for accountability. Users may be perceived committed not only to the cognitive effects that were originally intended: because of the mismatch between what is considered common ground by the audience, users would be perceived responsible for *all* the cognitive effects that were caused by their communicative stimulus. For example, while offensive jokes may be condoned[41] in settings in which certain contextual assumptions (moral values, audience's vulnerability, conversational habits) are salient and shared between communicator and audience, the same communicative behaviour would yield a different degree of responsibility if expressed in a way that can potentially be exposed to different audiences and thus generate different implications (endorsement of specific values, and so on).

Because of the intrinsic risk of such array of interpretations stemming from public communicative stimuli, such communicators would be hold particularly responsible for the kind of behaviour they manifest online; for the lack of preventive intentions manifested when posting controversial content online, and thus for contributing to create a digital environment that is florid for unchecked or fake news. As such, there is an assumption of 'epistemic authority' over those who publish online, which may be parasitic on the attitude we hold towards newspapers and sources who distribute information over audience that are not restricted to the one that was exposed to the production of the communicative act. As Szegőfi and Heintz (2022) propose, and surely for different reason that are related to the epistemic authority and reputation of such institutions, epistemic vigilance can be, and often is, deferred to newspapers and similar institutions. A commitment failure from the side of such institutions, i.e., the spread of fake news and inappropriate or offensive content, will raise an outrage that is a function of the degree of deferred epistemic vigilance that they carry.

Digital communication has a potential consequence that some users may change their course of action because of what is communicated online: they may misinterpret, get hurt, or spread sarcastic content as actual content. Exactly because online content has a potentially unrestricted outreach, the audience will hold them particularly responsible. Again, the higher

[41] The fact that they may be condoned does not entail that they should be condoned.

the likelihood that a partner, or an audience, will rely on what is communicated, the higher the perception of commitment and the cognitive responsibilities related with it.

References

Aboody, R., Yousif, S. R., Sheskin, M., & Keil, F. C. (2022). Says who? Children consider informants' sources when deciding whom to believe. *Journal of Experimental Psychology: General*, No Pagination Specified-No Pagination Specified. https://doi.org/10.1037/xge0001198

Adamic, L. A., Lento, T. M., Adar, E., & Ng, P. C. (2016). Information Evolution in Social Networks. *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, 473–482. https://doi.org/10.1145/2835776.2835827

Akdeniz, A., & Veelen, M. van. (2021). The evolution of morality and the role of commitment. *Evolutionary Human Sciences*, *3*. https://doi.org/10.1017/ehs.2021.36

Altay, S., Berriche, M., & Acerbi, A. (2021). *Misinformation on Misinformation: Conceptual and Methodological Challenges*. PsyArXiv. https://doi.org/10.31234/osf.io/edqc8

Altay, S., Hacquin, A.-S., & Mercier, H. (2020). Why do so few people share fake news? It hurts their reputation. *New Media & Society*, 1461444820969893. https://doi.org/10.1177/1461444820969893

Angel, K. (2022). *Tomorrow Sex Will Be Good Again: Women and Desire in the Age of Consent*. New Left Books Ltd. https://www.bloomsbury.com/au/tomorrow-sex-will-be-good-again-9781788739207/

Antomo, M., Müller, S., Paul, K., Paluch, M., & Thalmann, M. (2018). When children aren't more logical than adults: An empirical investigation of lying by falsely implicating. *Journal of Pragmatics*, *138*, 135–148. https://doi.org/10.1016/j.pragma.2018.09.010

Arden, R. (2020, October 3). Time and Perceptions of Trustworthiness—The Row over a Novel Study. *Quillette*. https://quillette.com/2020/10/03/time-and-perceptions-of-trustworthiness-the-row-over-a-novel-study/

Atherton, G., Sebanz, N., & Cross, L. (2019). Imagine All The Synchrony: The effects of actual and imagined synchronous walking on attitudes towards marginalised groups. *PLOS ONE*, *14*(5), e0216585. https://doi.org/10.1371/journal.pone.0216585

Austin, J. L. (1962). *How to do things with words*. Cambridge, MA: Harvard University Press.

Axelrod, R. (1984). *The Evolution of Cooperation*. Basic Books.

Bach, K., & Harnish, R. M. (1979). *Linguistic Communication and Speech Acts*. The MIT Press.

Back, I. H. (2010). Commitment bias: Mistaken partner selection or ancient wisdom? *Evolution and Human Behavior*, *31*(1), 22–28. https://doi.org/10.1016/j.evolhumbehav.2009.06.006

Barclay, P. (2016). Biological markets and the effects of partner choice on cooperation and friendship. *Current Opinion in Psychology*, *7*, 33–38. https://doi.org/10.1016/j.copsyc.2015.07.012

Barclay, P. (2017). Bidding to Commit. *Evolutionary Psychology*, *15*(1), 147470491769074. https://doi.org/10.1177/1474704917690740

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01.

Battigalli, P., & Dufwenberg, M. (2007). Guilt in Games. *The American Economic Review*, *97*(2), 170–176.

Baumard, N., André, J.-B., & Sperber, D. (2013). A mutualistic approach to morality: The evolution of fairness by partner choice. *Behavioral and Brain Sciences*, *36*(01), 59–78. https://doi.org/10.1017/S0140525X11002202

Belot, M., Bhaskar, V., & van de Ven, J. (2010). Promises and cooperation: Evidence from a TV game show. *Journal of Economic Behavior & Organization*, *73*(3), 396–405. https://doi.org/10.1016/J.JEBO.2010.01.001

Bentham, J. (2007). *An Introduction to the Principles of Morals and Legislation*. Dover Publications. (Original work published 1789)

Beres, M. A. (2007). 'Spontaneous' Sexual Consent: An Analysis of Sexual Consent Literature. *Feminism & Psychology*, *17*(1), 93–108. https://doi.org/10.1177/0959353507072914

Beres, M. A., Herold, E., & Maitland, S. B. (2004). Sexual Consent Behaviors in Same-Sex Relationships. *Archives of Sexual Behavior*, *33*, 475–486. https://doi.org/10.1023/B:ASEB.0000037428.41757.10

Berg, J., Dickhaut, J., & McCabe, K. A. (1995). Trust, Reciprocity, and Social History. *Games and Economic Behavior*, *10*(1), 122–142.

Bermúdez, J. L. (2005). *Philosophy of Psychology*. Routledge. https://www.taylorfrancis.com/books/mono/10.4324/9780203642405/philosophy-psychology-jose-luis-bermudez

Beyssade, C., & Marandin, J.-M. (2009). Commitment: Une attitude dialogique. *Langue Française*, *2*, 89–197. https://doi.org/10.3917/lf.162.0089

Bezuidenhout, A., & Cutting, J. C. (2002). Literal meaning, minimal propositions, and pragmatic processing. *Journal of Pragmatics*, *34*(4), 433–456. https://doi.org/10.1016/S0378-2166(01)00042-X

Bicchieri, C. (2006). *The Grammar of Society. The Nature and Dynamics of Social Norms*. Cambridge: Cambridge University Press.

Boisgontier, M. P., & Cheval, B. (2016). The anova to mixed model transition. *Neuroscience & Biobehavioral Reviews*, *68*, 1004–1005. https://doi.org/10.1016/j.neubiorev.2016.05.034

Bonalumi, F., Isella, M., & Michael, J. (2019). Cueing Implicit Commitment. *Review of Philosophy and Psychology*, *10*(4), 669–688. https://doi.org/10.1007/s13164-018-0425-0

Bonalumi, F., Mahr, J. B., Marie, P., & Pouscoulous, N. (2022). *Beyond the implicit/explicit dichotomy: The pragmatics of commitment, accountability, and plausible deniability*. https://doi.org/10.31234/osf.io/z2bqt

Bonalumi, F., Michael, J., & Heintz, C. (2021). Perceiving commitments: When we both know that you are counting on me. *Mind & Language*, mila.12333. https://doi.org/10.1111/mila.12333

Bonalumi, F., Scott-Phillips, T., Tacha, J., & Heintz, C. (2020). Commitment and Communication: Are we committed to what we mean, or what we say? *Language and Cognition*, 1–25. https://doi.org/10.1017/langcog.2020.2

Boogaart, R., Jansen, H., & van Leeuwen, M. (2020). "Those are Your Words, Not Mine!" Defence Strategies for Denying Speaker Commitment. *Argumentation*, *35*, 209–235. https://doi.org/10.1007/s10503-020-09521-3

Borg, E. (2017). Explanatory Roles for Minimal Content. *Nous*, *0*, 1–27. https://doi.org/10.1111/nous.12217

Boulat, K., & Maillat, D. (2017). She said you said I saw it with my own eyes: A pragmatic account of commitment. In J. Blochowiak, C. Grisot, S. Durrleman, & C. Länzingler (Eds.), *Formal Models in the Study of Language* (pp. 261–279). Springer.

Brandom, R. (1994). *Making It Explicit: Reasoning, Representing, and Discursive Commitment*. Harvard University Press.

Bratman, M. E. (1987). *Intention, Plans, and Practical Reason*. Harvard University Press.

Bratman, M. E. (1992). Shared Cooperative Activity. *Philosophical Psychology*, *101*(2), 327–341. https://doi.org/10.2307/2182828

Bronston v. United States, 409 U.S. No. 352 (1973). https://supreme.justia.com/cases/federal/us/409/352/

Brown, P., & Levinson, S. C. (1987). *Politeness: Some universals in language usage*. Cambridge University Press.

Camerer, C. F. (2003). *Behavioral Game Theory. Experiments in Strategic Interaction*. Princeton University Press. https://press.princeton.edu/books/hardcover/9780691090399/behavioral-game-theory

Carpenter, M., & Liebal, K. (2011). Joint attention, communication, and knowing together in infancy. In A. Seeman (Ed.), *Joint Attention: New Developments in Psychology, Philosophy of Mind, and Social Neuroscience* (pp. 159–182). Cambridge, MA: The MIT Press.

Carr, C. T., Schrock, D. B., & Dauterman, P. (2012). Speech Acts Within Facebook Status Messages. *Journal of Language and Social Psychology*, *31*(2), 176–196. https://doi.org/10.1177/0261927X12438535

Carson, T. L. (2006). The Definition of Lying. *Noûs*, *40*(2), 284–306. https://doi.org/10.1111/j.0029-4624.2006.00610.x

Carson, T. L. (2010). *Lying and Deception: Theory and Practice*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199577415.001.0001

Carston, R. (2002). *Thoughts and Utterances: The Pragmatics of Explicit Communication*. Blackwell.

Carston, R. (2004). Truth-Conditional Content and Conversational Implicature. In C. Bianchi (Ed.), *The Semantics/Pragmatics Distinction*. CSLI Publication.

Castelain, T., Bernard, S., & Mercier, H. (2018). Evidence that Two-Year-Old Children are Sensitive to Information Presented in Arguments. *Infancy*, *23*(1), 124–135.

Castelain, T., Bernard, S., Van der Henst, J.-B., & Mercier, H. (2016). The influence of power and reason on young Maya children's endorsement of testimony. *Developmental Science*, *19*(6), 957–966. https://doi.org/10.1111/desc.12336

Çelik, B., Ergut, N., & Allen, J. W. P. (2022). *Relying on Evidentials: The Role of Evidentiality as a Cue for Children's Reliability Judgments.* [Manuscript in preparation].

Charness, G., & Dufwenberg, M. (2006). Promises and Partnership. *Econometrica*, *74*(6), 1579–1601.

Charness, G., & Rabin, M. (2010). Understanding Social Preferences with Simple Tests. *The Quarterly Journal of Economics*, *117*(3), 817–869. https://doi.org/10.2307/4132490

Chennells, M., & Michael, J. (2018). Effort and performance in a cooperative activity are boosted by perception of a partner's effort. *Nature*, *8*(15692). https://doi.org/10.1038/s41598-018-34096-1

Chennells, M., & Michael, J. (2021, July 12). *Ask not and ye shall receive: When explicit agreements undermine the sense of commitment*. [Talk]. Speakers' commitment to their utterance content and hearers' epistemic vigilance in accepting that content. UCL Workshop., London. https://www.lahp.ac.uk/event/speakers-commitment-to-their-utterance-content-and-hearers-epistemic-vigilance-in-accepting-that-content-previous-staff-led-events/

Chennells, M., & Michael, J. (2022). Breaking the right way: A closer look at how we dissolve commitments. *Phenomenology and the Cognitive Sciences*. https://doi.org/10.1007/s11097-022-09805-x

Chennells, M., Woźniak, M., Butterfill, S., & Michael, J. (2022). *Coordinated Decision-Making Boosts Altruistic Motivation – But Not Trust* [Manuscript in preparation].

Christensen, R. H. B. (2019). *ordinal—Regression Models for Ordinal Data*. (R package version 2019.12-10) [Computer software]. https://CRAN.R-project.org/package=ordinal

Clark, B. (2013). *Relevance Theory*. Oxford University Press.

Clark, H. H. (2006). Social actions, social commitments. In N. J. Enfield & S. C. Levinson (Eds.), *Roots of Human Sociality: Culture, Cognition, and Human Interaction* (pp. 126–150). Bloomsbury.

Corriveau, K. H., & Kurkul, K. E. (2014). "Why Does Rain Fall?": Children Prefer to Learn From an Informant Who Uses Noncircular Explanations. *Child Development*, *85*(5), 1827–1835. https://doi.org/10.1111/cdev.12240

Cross, L., Wilson, A. D., & Golonka, S. (2016). How Moving Together Brings Us Together: When Coordinated Rhythmic Movement Affects Cooperation. *Frontiers in Psychology*, *7*. https://www.frontiersin.org/article/10.3389/fpsyg.2016.01983

Csibra, G., & Gergely, G. (2009). Natural pedagogy. *Trends in Cognitive Sciences*, *13*(4), 127–132. https://doi.org/10.1016/j.tics.2009.01.005

Dana, J., Cain, D. M., & Dawes, R. M. (2006). What you don't know won't hurt me: Costly (but quiet) exit in dictator games. *Organizational Behavior and Human Decision Processes*, *100*(2), 193–201. https://doi.org/10.1016/j.obhdp.2005.10.001

Dana, J., Weber, R. A., & Kuang, J. X. (2007). Exploiting moral wiggle room: Experiments demonstrating an illusory preference for fairness. *Economic Theory*, *33*(1), 67–80. https://doi.org/10.1007/s00199-006-0153-z

Danovitch, J. H., & Lane, J. D. (2020). Children's belief in purported events: When claims reference hearsay, books, or the internet. *Journal of Experimental Child Psychology*, *193*, 104808. https://doi.org/10.1016/j.jecp.2020.104808

Danziger, E. (2010). On trying and lying: Cultural configurations of Grice's Maxim of Quality. *Intercultural Pragmatics*, *7*(2), 199–219. https://doi.org/10.1515/iprg.2010.010

Darwall, S. L. (2006). *The second-person standpoint: Morality, respect, and accountability*. Cambridge, MA: Harvard University Press.

De Brabanter, P., & Dendale, P. (2008). Commitment: The term and the notions. *Belgian Journal of Linguistics*, *22*, 1–14. https://doi.org/10.1075/bjl.22.01de

Degen, J., Trotzke, A., Scontras, G., Wittenberg, E., & Goodman, N. D. (2019). Definitely, maybe: A new experimental paradigm for investigating the pragmatics of evidential devices across languages. *Journal of Pragmatics*, *140*, 33–48. https://doi.org/10.1016/j.pragma.2018.11.015

Dennett, D. C. (1987). *The Intentional Stance*. The MIT Press. https://mitpress.mit.edu/books/intentional-stance

Dinges, A., & Zakkou, J. (2021). *On Deniability* [Manuscript in preparation].

Dobson, A. J., & Barnett, A. G. (2008). *An introduction to generalized linear models* (3rd ed.). New York, NY: Chapman & Hall/CRC.

Domberg, A., Köymen, B., & Tomasello, M. (2018). Children's reasoning with peers in cooperative and competitive contexts. *British Journal of Developmental Psychology*, *36*, 64–77. https://doi.org/10.1111/bjdp.12213

Doran, R., Baker, R., McNabb, Y., Larson, M., & Ward, G. (2009). On the Non-Unified Nature of Scalar Implicature: An Empirical Investigation. *International Review of Pragmatics*, *1*, 1–38. https://doi.org/10.1163/187730909X12538045489854

Doran, R., Ward, G., Larson, M., McNabb, Y., & Baker, R. E. (2012). A Novel Experimental Paradigm for Distinguishing Between 'What is Said' and 'What is Implicated'. *Language*, *88*(1), 124–154. https://doi.org/10.1353/lan.2012.0008

Dripps, D. (2009). For a Negative, Normative Model of Consent, With a Comment on Preference-Skepticism. *Legal Theory*, *2*(2), 113–120. https://doi.org/10.1017/S1352325200000422

Dufwenberg, M., & Gneezy, U. (2000). Measuring Beliefs in an Experimental Lost Wallet Game. *Games and Economic Behavior*, *30*(2), 163–182. https://doi.org/10.1006/game.1999.0715

Ekman, P., Davidson, R. J., & Friesen, W. V. (1990). The Duchenne smile: Emotional expression and brain physiology: II. *Journal of Personality and Social Psychology*, *58*(2), 342–353.

Elfenbein, H. A., & Ambady, N. (2002). On the universality and cultural specificity of emotion recognition: A meta-analysis. *Psychological Bulletin*, *128*(2), 203–235. https://doi.org/10.1037/0033-2909.128.2.203

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Fehr, E., & Gächter, S. (2000). Cooperation and punishment in public goods experiments. *American Economic Review*, *90*(4), 980–994.

Fessler, D. M. T., & Quintelier, K. (2014). Suicide Bombers, Weddings, and Prison Tattoos. An Evolutionary Perspective on Subjective Commitment and Objective Commitment. In R. Joyce, K. Sterelny, & B. Calcott (Eds.), *Cooperation and its evolution* (pp. 459–484). The MIT Press.

Fiske, A. P. (1992). The four elementary forms of sociality: Framework for a unified theory of social relations. *Psychological Review*, *99*(4), 689–723.

Fiske, S. T., Cuddy, A. J. C., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences*, *11*(2), 77–83. https://doi.org/10.1016/j.tics.2006.11.005

Fitneva, S. A. (2008). The role of evidentiality in Bulgarian children's reliability judgments*. *Journal of Child Language*, *35*(4), 845–868. https://doi.org/10.1017/S0305000908008799

Forstmeier, W., & Schielzeth, H. (2011). Cryptic multiple hypotheses testing in linear models: Overestimated effect sizes and the winner's curse. *Behavioral Ecology and Sociobiology*, *65*(1), 47–55. https://doi.org/10.1007/s00265-010-1038-5

Frank, R. H. (1988). *Passion within reason*. W.W. Norton & Company.

Franke, M., Dulcinati, G., & Pouscoulous, N. (2020). Strategies of Deception: Under-Informativity, Uninformativity, and Lies-Misleading With Different Kinds of Implicature. *Topics in Cognitive Science*, *12*(2), 583–607. https://doi.org/10.1111/tops.12456

Gallese, V., & Goldman, A. (1998). Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Sciences*, *2*(12), 493–501. https://doi.org/10.1016/S1364-6613(98)01262-5

Geurts, B. (2019). Communication as commitment making: Speech acts, implicatures, common ground. *Theoretical Linguistics*, *45*(1–2), 1–30. https://doi.org/10.1515/tl-2019-0001

Gibbs, R. W., & Moise, J. F. (1997). Pragmatics in understanding what is said. *Cognition*, *62*(1), 51–74. https://doi.org/10.1016/S0010-0277(96)00724-X

Gilbert, M. (1990). Walking Together: A paradigmatic social phenomenon. *Midwest Studies in Philosophy*, *15*, 1–14. https://doi.org/10.1111/j.1475-4975.1990.tb00202.x

Gilbert, M. (2006). Rationality in collective action. *Philosophy of the Social Sciences*, *36*(1), 3–17.

Gilbert, M. (2009). Shared intention and personal intention. *Philosophical Studies*, *144*, 167–187.

Gilbert, M. (2014). *Joint commitment. How we make the social world*. New York, NY: Oxford University Press.

Gilbert, M. (2017). Joint commitment. In M. Janković & K. Ludwig (Eds.), *The Routledge Handbook of Collective Intentionality* (pp. 130–139). New York, NY: Routledge.

Gomez-Lavin, J., & Rachar, M. (2019). Normativity in joint action. *Mind & Language*, *34*(1), 97–120. https://doi.org/10.1111/mila.12195

Gräfenhain, M., Behne, T., Carpenter, M., & Tomasello, M. (2009). Young children's understanding of joint commitments. *Developmental Psychology*, *45*(5), 1430–1443. https://doi.org/10.1037/a0016122

Gräfenhain, M., Carpenter, M., & Tomasello, M. (2013). Three-year-olds' understanding of the consequences of joint commitments. *PLoS ONE*, *8*(9). https://doi.org/10.1371/journal.pone.0073039

Green, A., Siposova, B., Kita, S., & Michael, J. (2021). Stopping at nothing: Two-year-olds differentiate between interrupted and abandoned goals. *Journal of Experimental Child Psychology*, *209*, 105171. https://doi.org/10.1016/j.jecp.2021.105171

Grice, H. P. (1957). Logic and Conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics. Vol. 3: Speech Acts* (pp. 41–58). Academic Press.

Grice, H. P. (1989). *Studies in the Way of Words.* Harvard University Press.

Grueneisen, S., & Tomasello, M. (2020). The development of coordination via joint expectations for shared benefits. *Developmental Psychology*, *56*(6), 1149–1156. https://doi.org/10.1037/dev0000936

Guala, F. (2012). Reciprocity: Weak or strong? What punishment experiments do (and do not) demonstrate. *Behavioral and Brain Sciences*, *35*(01), 1–15. https://doi.org/10.1017/S0140525X11000069

Guala, F., & Mittone, L. (2010). How history and convention create norms: An experimental study. *Journal of Economic Psychology*, *31*(4), 749–756. https://doi.org/10.1016/j.joep.2010.05.009

Guercio, N. L., & Caso, R. (2021). *An Account of Overt Intentional Dogwhistling* [Manuscript in preparation].

Guess, A., Nagler, J., & Tucker, J. (2019). Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science Advances*, *5*(1), eaau4586. https://doi.org/10.1126/sciadv.aau4586

Güth, W., Schmittberger, R., & Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior & Organization*, *3*(4), 367–388. https://doi.org/10.1016/0167-2681(82)90011-7

Hamann, K., Warneken, F., & Tomasello, M. (2012). Children's developing commitments to joint goals. *Child Development*, *83*(1), 137–145. https://doi.org/10.1111/j.1467-8624.2011.01695.x

Hamblin, C. L. (1971). Mathematical models of dialogue 1. *Theoria*, *37*(2).

Hamilton, W. D. (1964). The genetical evolution of social behaviour. I. *Journal of Theoretical Biology*, *7*(1), 1–16. https://doi.org/10.1016/0022-5193(64)90038-4

Hardin, K. J. (2010). The Spanish notion of Lie: Revisiting Coleman and Kay. *Journal of Pragmatics*, *42*(12), 3199–3213. https://doi.org/10.1016/j.pragma.2010.07.006

Heintz, C., Celse, J., Giardini, F., & Max, S. (2015). Facing expectations: Those that we prefer to fulfil and those that we disregard. *Judgment and Decision Making*, *10*(5), 442–455.

Heintz, C., Karabegović, M., & Molnar, A. (2016). The co-evolution of honesty and strategic vigilance. *Frontiers in Psychology*, *7*, Article 1503. https://doi.org/10.3389/fpsyg.2016.01503

Heintz, C., & Scott-Phillips, T. (in press). *Expression Unleashed*. PsyArXiv. https://doi.org/10.31234/osf.io/mcv5b

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *The Behavioral and Brain Sciences*, *33*(2–3), 61–83; discussion 83-135. https://doi.org/10.1017/S0140525X0999152X

Hirshleifer, J. (2001). On the Emotions as Guarantors of Threats and Promises. In *The Dark Side of the Force: Economic Foundations of Conflict Theory* (pp. 198–219). Cambridge University Press. http://econpapers.repec.org/RePEc:cla:uclawp:337

Homer. (1919). *The Odyssey* (A. T. Murray, Trans.). W. Heinemann; G.P. Putnam's Sons. (Original work published 800 B.C.E.)

Horn, L. R. (1984). Towards a new taxonomy for pragmatic inference: Q- and R-based implicature. Meaning, form, and use in context. In D. Schiffrin (Ed.), *Meaning, Form, and Use in Context: Linguistic Applications* (pp. 11–42). Georgetown University Press.

H.R. 5430—116th Congress: United States-Mexico-Canada Agreement Implementation Act., (2022). Retrieved from https://www.govtrack.us/congress/bills/116/hr5430

Hruschka, D. (2020). *Cultural Diversity in the Meaning of Lies, Deceptions, and other Misrepresentations*. SocArXiv. https://doi.org/10.31235/osf.io/8puwc

Hume, D. (2000). *A Treatise of Human Nature* (D. F. Norton & M. J. Norton, Eds.). Oxford University Press. (Original work published 1739–1740)

Isella, M., Kanngiesser, P., & Tomasello, M. (2019). Children's selective trust in promises. *Child Development*, *90*(6), e868–e887. https://doi.org/10.1111/cdev.13105

Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in Cognitive Sciences*, *20*(8), 589–604. https://doi.org/10.1016/j.tics.2016.05.011

Kachel, U., Svetlova, M., & Tomasello, M. (2018). Three-year-olds' reactions to a partner's failure to perform her role in a joint commitment. *Child Development*, *89*(5), 1691–1703. https://doi.org/10.1111/cdev.12816

Kachel, U., Svetlova, M., & Tomasello, M. (2019). Three- and 5-year-old children's understanding of how to dissolve a joint commitment. *Journal of Experimental Child Psychology*, *184*, 34–47. https://doi.org/10.1016/j.jecp.2019.03.008

Kachel, U., & Tomasello, M. (2019). 3- and 5-year-old children's adherence to explicit and implicit joint commitments. *Developmental Psychology*. https://doi.org/10.1037/dev0000632

Kanngiesser, P., Mammen, M., & Tomasello, M. (2021). Young children's understanding of justifications for breaking a promise. *Cognitive Development*, *60*, 101127. https://doi.org/10.1016/j.cogdev.2021.101127

Kessler, G. (2019, January 9). A history of Trump's promises that Mexico would pay for the wall, which it refuses to do. *The Washington Post*. https://www.washingtonpost.com/politics/2019/live-updates/trump-white-house/live-fact-checking-and-analysis-of-president-trumps-immigration-speech/a-history-of-trumps-promises-that-mexico-would-pay-for-the-wall-which-it-refuses-to-do/

Kitzinger, C., & Frith, H. (1999). Just Say No? The Use of Conversation Analysis in Developing a Feminist Perspective on Sexual Refusal. *Discourse & Society*, *10*(3), 293–316. https://doi.org/10.1177/0957926599010003002

Klein, B. (2019, January 10). Trump claims 'obviously' Mexico isn't going to write a check for a border wall. *CNN*. https://www.cnn.com/2019/01/10/politics/trump-mexico-pay-wall/index.html

Koenig, M. A. (2012). Beyond Semantic Accuracy: Preschoolers Evaluate a Speaker's Reasons: Children's Understanding of Reasons. *Child Development*, *83*(3), 1051–1063. https://doi.org/10.1111/j.1467-8624.2012.01742.x

Koenig, M. A., Clément, F., & Harris, P. L. (2004). Trust in Testimony: Children's Use of True and False Statements. *Psychological Science*, *15*(10), 694–698. https://doi.org/10.1111/j.0956-7976.2004.00742.x

Koenig, M. A., & Harris, P. L. (2005). Preschoolers Mistrust Ignorant and Inaccurate Speakers. *Child Development*, *76*(6), 1261–1277. https://doi.org/10.1111/j.1467-8624.2005.00849.x

Kokal, I., Engel, A., Kirschner, S., & Keysers, C. (2011). Synchronized Drumming Enhances Activity in the Caudate and Facilitates Prosocial Commitment—If the Rhythm Comes Easily. *PLOS ONE*, *6*(11), e27272. https://doi.org/10.1371/journal.pone.0027272

Koomen, R., Grueneisen, S., & Herrmann, E. (2020). Children delay gratification for cooperative ends. *Psychological Science*, *31*(2), 139–148. https://doi.org/10.1177/0956797619894205

Kruschke, J. K. (2015). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan.* (2nd ed.). Academic Press.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*, 1–26. https://doi.org/10.18637/jss.v082.i13

Lane, J. D., Ronfard, S., & El-Sherif, D. (2018). The Influence of First-Hand Testimony and Hearsay on Children's Belief in the Improbable. *Child Development*, *89*(4), 1133–1140. https://doi.org/10.1111/cdev.12815

Launay, J., Dean, R. T., & Bailes, F. (2013). Synchronization can influence trust following virtual interaction. *Experimental Psychology*, *60*(1), 53–63. https://doi.org/10.1027/1618-3169/a000173

Lee, J. J., & Pinker, S. (2010). Rationales for Indirect Speech: The Theory of the Strategic Speaker. *Psychological Review*, *117*(3), 785–807. https://doi.org/10.1037/a0019688

Lehmann, E. L. (2006). *Nonparametrics. Statistical Methods Based on Ranks*. New York, NY: Springer-Verlag.

Leslie, A. M., Friedman, O., & German, T. P. (2004). Core mechanisms in 'theory of mind'. *Trends in Cognitive Sciences*, *8*(12), 528–533. https://doi.org/10.1016/j.tics.2004.10.001

Levinson, S. C. (2000). *Presumptive Meanings. The Theory of Generalized Conversational Implicature.* The MIT Press.

Levinson, S. C. (2006). On the human 'interaction engine' | Max Planck Institute. In N. J. Enfield & S. C. Levinson (Eds.), *Roots of Human Sociality: Culture, Cognition and Interaction* (pp. 39–69). Berg Publishers. https://www.mpi.nl/publications/item59332/human-interaction-engine

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

Lewis, D. K. (1978). Truth in Fiction. *American Philosophical Quarterly*, *15*, 37–46.

Li, P. H., & Koenig, M. A. (2020). Children's Evaluations of Informants and their Surprising Claims in Direct and Overheard Contexts. *Journal of Cognition and Development*, *21*(3), 425–446. https://doi.org/10.1080/15248372.2020.1745208

Liddell, T. M., & Kruschke, J. K. (2018). Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology*, *79*, 328–348. https://doi.org/10.1016/j.jesp.2018.08.009

Loftus, E. F. (1981). Eyewitness Testimony: Psychological Research and Legal Thought. *Crime and Justice*, *3*, 105–151. https://doi.org/10.1086/449078

Löhr, G. (2021). Commitment engineering: Conceptual engineering without representations. *Synthese*, *199*(5), 13035–13052. https://doi.org/10.1007/s11229-021-03365-4

Luke, S. G. (2017). Evaluating significance in linear mixed-effects models in R. *Behavior Research Methods*, *49*(4), 1494–1502. https://doi.org/10.3758/s13428-016-0809-y

MacCormick, N., & Raz, J. (1972). Voluntary Obligations and Normative Powers. *Proceedings of the Aristotelian Society, Supplementary Volumes*, *46*(1972), 59–102. https://doi.org/10.1097/EDE.ObO13e31812e5535

Mahr, J. B., & Csibra, G. (2018). Why do we remember? The communicative function of episodic memory. *Behavioral and Brain Sciences*, *41*, 1–16. https://doi.org/10.1017/S0140525X17000012

Mahr, J. B., & Csibra, G. (2020). Witnessing, Remembering, and Testifying: Why the Past Is Special for Human Beings. *Perspectives on Psychological Science*, *15*(2), 428–443. https://doi.org/10.1177/1745691619879167

Mahr, J. B., & Csibra, G. (2021). The effect of source claims on statement believability and speaker accountability. *Memory & Cognition*, *49*(8), 1505–1525. https://doi.org/10.3758/s13421-021-01186-x

Mant, C. M., & Perner, J. (1988). The child's understanding of commitment. *Developmental Psychology*, *24*(3), 343–351. https://doi.org/10.1037/0012-1649.24.3.343

Marsili, N. (2016). Lying by Promising. A study on insincere illocutionary acts. *International Review of Pragmatics*, *8*(2), 271–313.

Marsili, N. (2021a). Retweeting: Its linguistic and epistemic value. *Synthese*, *198*(11), 10457–10483. https://doi.org/10.1007/s11229-020-02731-y

Marsili, N. (2021b). Lying, speech acts, and commitment. *Synthese*, *199*(1), 3245–3269. https://doi.org/10.1007/s11229-020-02933-4

Mascaro, O., & Kovács, Á. (2022). The Origins of Trust: Humans' Reliance on Communicative Cues Supersedes Firsthand Experience During the Second Year of Life. *Developmental Science*, e13223. https://doi.org/10.1111/desc.13223

Maynard Smith, J. (1982). *Evolution and the Theory of Games*. Cambridge University Press. https://www.cambridge.org/at/academic/subjects/life-sciences/evolutionary-biology/evolution-and-theory-games?format=PB&isbn=9780521288842

Mazzarella, D. (2021). "I didn't mean to suggest anything like that!": Deniability and context reconstruction. *Mind & Language*, 1–19. https://doi.org/10.1111/mila.12377

Mazzarella, D., Reinecke, R., Noveck, I. A., & Mercier, H. (2018). Saying, presupposing and implicating: How pragmatics modulates commitment. *Journal of Pragmatics*, *133*(August), 15–27. https://doi.org/10.1016/j.pragma.2018.05.009

McEllin, L., Felber, A., & Michael, J. (2022). The Fruits of our Labour: Interpersonal coordination generates commitment by signaling a willingness to adapt. *Quarterly Journal of Experimental Psychology*, 17470218221079830. https://doi.org/10.1177/17470218221079830

Meibauer, J. (2018). The Linguistics of Lying. *Annu. Rev. Linguist.*, *4*, 357–377.

Mercier, H. (2017). How Gullible are We? A Review of the Evidence from Psychology and Social Science. *Review of General Psychology*, *21*(2), 103–122. https://doi.org/10.1037/gpr0000111

Mercier, H., Bernard, S., & Clément, F. (2014). Early sensitivity to arguments: How preschoolers weight circular arguments. *Journal of Experimental Child Psychology*. https://doi.org/10.1016/j.jecp.2013.11.011

Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *The Behavioral and Brain Sciences*, *34*(2), 57–74; discussion 74-111. https://doi.org/10.1017/S0140525X10000968

Mercier, H., Sudo, M., Castelain, T., Bernard, S., & Matsui, T. (2017). Japanese preschoolers' evaluation of circular and non-circular arguments. *European Journal of Developmental Psychology*, *15*(5), 493–505. https://doi.org/10.1080/17405629.2017.1308250

Michael, J. (2022). *The Philosophy and Psychology of Commitment*. Routledge/Taylor & Francis Group.

Michael, J., & Pacherie, E. (2015). On commitments and other uncertainty reduction tools in joint action. *Journal of Social Ontology*, *1*(1). https://doi.org/10.1515/jso-2014-0021

Michael, J., Sebanz, N., & Knoblich, G. (2016a). The sense of commitment: A minimal approach. *Frontiers in Psychology*, *6*, 1968. https://doi.org/10.3389/fpsyg.2015.01968

Michael, J., Sebanz, N., & Knoblich, G. (2016b). Observing joint action: Coordination creates commitment. *Cognition*, *157*, 106–113. https://doi.org/10.1016/j.cognition.2016.08.024

Michael, J., & Székely, M. (2018). The developmental origins of commitment. *Journal of Social Philosophy*. https://doi.org/10.1016/bs.acdb.2017.10.006

Mill, J. S. (2014). *Utilitarianism*. Cambridge University Press. (Original work published 1863)

Misyak, J. B., & Chater, N. (2014). Virtual bargaining: A theory of social decision-making. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *369*(1655), 20130487. https://doi.org/10.1098/rstb.2013.0487

Moore, A. (2021, June 15). The sexual assault of sleeping women: The hidden, horrifying rape crisis in our bedrooms. *The Guardian*. https://www.theguardian.com/society/2021/jun/15/the-sexual-assault-of-sleeping-women-the-hidden-horrifying-crisis-in-britains-bedrooms

Moore, G. E. (2004). *Principia Ethica*. Dover Publications. (Original work published 1903)

Morency, P., Oswald, S., & De Saussure, L. (2008). Explicitness, Implicitness and Commitment Attribution: A Cognitive Pragmatic Approach. *Belgian Journal of Linguistics*, *22*(1), 197–219. https://doi.org/10.1075/bjl.22.10mor

Möschler, J. (2013). Is a speaker-based pragmatics possible? Or how can a hearer infer a speaker's commitment? *Journal of Pragmatics*, *48*(1), 84–97. https://doi.org/10.1016/j.pragma.2012.11.019

Nesse, R. M. (Ed.). (2001). *Evolution and the capacity for commitment*. Russell Sage Foundation.

Newman, N., Fletcher, R., Kalogeropoulos, A., Levy, D. A. L., & Nielsen, R. K. (2018). *Reuters Institute Digital News Report 2018*. Reuters Institute for the Study of Journalism. https://reutersinstitute.politics.ox.ac.uk/sites/default/files/digital-news-report-2018.pdf

Nicolle, S., & Clark, B. (1999). Experimental pragmatics and what is said: A response to Gibbs and Moise. *Cognition*, *69*(3), 337–354. https://doi.org/10.1016/S0010-0277(98)00070-5

Nielsen, M., Haun, D., Kärtner, J., & Legare, C. H. (2017). The persistent sampling bias in developmental psychology: A call to action. *Journal of Experimental Child Psychology*, *162*, 31–38. https://doi.org/10.1016/j.jecp.2017.04.017

Noë, R., & Hammerstein, P. (1994). Biological Markets: Supply and Demand Determine the Effect of Partner Choice in Cooperation, Mutualism and Mating. *Behavioral Ecology and Sociobiology*, *35*(1), 1–11.

Noveck, I. A. (2004). Pragmatic Inferences Related to Logical Terms. In I. A. Noveck & D. Sperber (Eds.), *Experimental Pragmatics. Palgrave Studies in Pragmatics, Language and Cognition.* (pp. 301–321). Palgrave Macmillan UK. https://doi.org/10.1057/9780230524125_14

Noveck, I. A., & Posada, A. (2003). Characterizing the time course of an implicature: An evoked potentials study. *Brain and Language*, *85*(2), 203–210. https://doi.org/10.1016/S0093-934X(03)00053-1

Noveck, I. A., & Sperber, D. (2007). The why and how of experimental pragmatics: The case of 'scalar inferences'. In *Advances in Pragmatics* (pp. 184–212). Palgrave.

Nowak, M. A., & Sigmund, K. (2005). Evolution of indirect reciprocity. *Nature*, *437*(7063), 1291–1298. https://doi.org/10.1038/nature04131

O'Byrne, R., Rapley, M., & Hansen, S. (2006). 'You Couldn't Say "No", Could You?': Young Men's Understandings of Sexual Refusal. *Feminism & Psychology*, *16*(2), 133–154. https://doi.org/10.1177/0959-353506062970

Ockenfels, A., & Werner, P. (2012). "Hiding behind a small cake" in a newspaper dictator game. *Journal of Economic Behavior & Organization*, *82*(1), 82–85.

Or, S., Ariel, M., & Peleg, O. (2017). The case of literally true propositions with false implicatures. In I. Chiluwa (Ed.), *Deception & Deceptive Communication. Motivations, Recognition Techniques and Behavioral Control.* Nova Science.

Ozturk, O., & Papafragou, A. (2016). The Acquisition of Evidentiality and Source Monitoring. *Language Learning and Development*, *12*(2), 199–230. https://doi.org/10.1080/15475441.2015.1024834

Palan, S., & Schitter, C. (2018). Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, *17*, 22–27. https://doi.org/10.1016/j.jbef.2017.12.004

Pineau, L. (1989). Date rape: A feminist analysis. *Law and Philosophy*, *8*(2), 217–243. https://doi.org/10.1007/BF00160012

Pinker, S. (2004). *The Language Instinct: How the Mind Creates Language* (Reprint edition). Morrow.

Pinker, S. (2007). The evolutionary social psychology of off-record indirect speech acts. *Intercultural Pragmatics*, *4*(4), 437–461. https://doi.org/10.1515/IP.2007.023

Pinker, S., Nowak, M. A., & Lee, J. J. (2008). The logic of indirect speech. *Proceedings of the National Academy of Sciences*, *105*(3), 833–838. https://doi.org/10.1073/pnas.0707192105

Plummer, M., Stukalov, A., & Denwood, M. (2021). *rjags—Bayesian Graphical Models using MCMC*. (R package version 4-12.) [Computer software]. https://cran.r-project.org/web/packages/rjags/rjags.pdf

R Core Team. (2020). *R: A language and environment for statistical computing. R Foundation for Statistical Computing*. https://www.R-project.org/

Raftery, B. (2014, September 14). IMDb Voters Are Tanking Indies Before They're Even Released. *WIRED*. https://www.wired.com/2016/09/imdb-voters-tanking-indies-theyre-even-released/

Rakoczy, H., Warneken, F., & Tomasello, M. (2007). 'This way!', 'No! That way!'-3-year olds know that two people can have mutually incompatible desires. *Cognitive Development*, *22*(1), 47–68. https://doi.org/10.1016/j.cogdev.2006.08.002

Rakoczy, H., Warneken, F., & Tomasello, M. (2008). The sources of normativity: Young children's awareness of the normative structure of games. *Developmental Psychology*, *44*(3), 875–881. https://doi.org/10.1037/0012-1649.44.3.875

Rawls, J. (1955). Two Concepts of Rules. *The Philosophical Review*, *64*(1), 3–32.

Rawls, J. (1971). *A theory of justice*. Harvard University Press.

Reboul, A. (2017). Is implicit communication a way to escape epistemic vigilance? In S. Assimakopoulos (Ed.), *Pragmatics at its Interfaces* (pp. 91–112). de Gruyter Mouton.

Récanati, F. (1993). *Direct Reference: From Language to Thought*. Blackwell.

Récanati, F. (2001). What is Said. *Synthese*, *128*(1–2), 75–91. https://doi.org/10.1023/A:1010383405105

Récanati, F. (2004). *Literal Meaning*. Cambridge University Press.

Reins, L. M., & Wiegmann, A. (2021). Is Lying Bound to Commitment? Empirically Investigating Deceptive Presuppositions, Implicatures, and Actions. *Cognitive Science*, *45*(2), e12936. https://doi.org/10.1111/cogs.12936

Roberts, G. (2005). Cooperation through interdependence. *Animal Behaviour*, *70*(4), 901–908. https://doi.org/10.1016/j.anbehav.2005.02.006

Robinson, E. J., & Whitcombe, E. L. (2003). Children's Suggestibility in Relation to their Understanding about Sources of Knowledge. *Child Development*, *74*(1), 48–62. https://doi.org/10.1111/1467-8624.t01-1-00520

Ronfard, S., Nelson, L., Dunham, Y., & Blake, P. R. (2019). How children use accuracy information to infer informant intentions and to make reward decisions. *Journal of Experimental Child Psychology*, *177*, 100–118. https://doi.org/10.1016/j.jecp.2018.07.017

Ronson, J. (2015, February 12). How One Stupid Tweet Blew Up Justine Sacco's Life. *The New York Times Magazine*. https://www.nytimes.com/2015/02/15/magazine/how-one-stupid-tweet-ruined-justine-saccos-life.html

RStudio Team. (2020). *RStudio: Integrated Development for R*. RStudio, PBC. http://www.rstudio.com/

Rusch, H., & Lütge, C. (2016). Spillovers from coordination to cooperation: Evidence for the interdependence hypothesis? *Evolutionary Behavioral Sciences*, *10*(4), 284–296.

Safra, L., Chevallier, C., Grèzes, J., & Baumard, N. (2020). Tracking historical changes in trustworthiness using machine learning analyses of facial cues in paintings. *Nature Communications*, *11*(1), 4728. https://doi.org/10.1038/s41467-020-18566-7

Santre, S. (2022). Cyberbullying in adolescents: A literature review. *International Journal of Adolescent Medicine and Health*. https://doi.org/10.1515/ijamh-2021-0133

Saul, J. M. (2012). *Lying, Misleading, and What Is Said*. Oxford University Press.

Scanlon, T. M. (1998). *What We Owe to Each Other*. Cambridge, MA: Harvard University Press.

Schelling, T. C. (1980). *The Strategy of Conflict*. Cambridge, MA: Harvard University Press.

Schiffer, T. (1972). *Meaning*. Cambridge, MA: Harvard University Press.

Schmidt, M. F. H., Butler, L. P., Heinz, J., & Tomasello, M. (2016). Young Children See a Single Action and Infer a Social Norm: Promiscuous Normativity in 3-Year-Olds. *Psychological Science*. https://doi.org/10.1177/0956797616661182

Schrift, R. Y., & Parker, J. R. (2014). Staying the Course: The Option of Doing Nothing and Its Impact on Postchoice Persistence. *Psychological Science*, *25*(3), 772–780. https://doi.org/10.1177/0956797613516801

Scott-Phillips, T. (2015). *Speaking our minds: Why human communication is different, and how language evolved to make it special*. Macmillan International Higher Education.

Searle, J. R. (1969). *Speech Acts: An essay in the philosophy of language*. Cambridge: Cambridge University Press.

Searle, J. R. (1990). Collective Intentions and Actions. In P. R. Cohen, J. Morgan, & M. E. Pollack (Eds.), *Intentions in Communication* (pp. 401–416). Bradford Books, MIT Press.

Searle, J. R. (2010). *Making the Social World*. New York, NY: Oxford University Press.

Sebanz, N., Bekkering, H., & Knoblich, G. (2006). Joint action: Bodies and minds moving together. *Trends in Cognitive Sciences*, *10*(2), 70–76. https://doi.org/10.1016/j.tics.2005.12.009

Shpall, S. (2014). Moral and rational commitment. *Philosophy and Phenomenological Research*, *88*(1), 146–172. https://doi.org/10.1111/j.1933-1592.2012.00618.x

Sidgwick, H. (2011). *The methods of ethics*. Cambridge University Press. (Original work published 1874)

Singmann, H., Bolker, B., Westfall, J., Aust, F., & Ben-Shachar, M. S. (2021). *afex: Analysis of Factorial Experiments*. (R package version 0.28-1.) [Computer software]. https://CRAN.R-project.org/package=afex

Siposova, B., Tomasello, M., & Carpenter, M. (2018). Communicative eye contact signals a commitment to cooperate for young children. *Cognition*, *179*, 192–201. https://doi.org/10.1016/j.cognition.2018.06.010

Smith, A. (2006). *The theory of moral sentiments*. Dover Publications. (Original work published 1759)

Soltys, J., Terkourafi, M., & Katsos, N. (2014). Disentangling Politeness Theory and the Strategic Speaker approach: Theoretical considerations and empirical predictions. *Intercultural Pragmatics*, *11*(1), 31–56. https://doi.org/10.1515/ip-2014-0002

Sommers, R. (2019). Commonsense Consent. *Yale Law Journal*, *129*(8), 2232–2325.

Song, R., Over, H., & Carpenter, M. (2016). Young children discriminate genuine from fake smiles and expect people displaying genuine smiles to be more prosocial. *Evolution and Human Behavior*, *37*(6), 490–501. https://doi.org/10.1016/j.evolhumbehav.2016.05.002

Sperber, D. (1994). How do we communicate? In J. Brockman & K. Matson (Eds.), *How things are: A science toolkit for the mind* (pp. 191–199). Morrow.

Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic vigilance. *Mind and Language*, *25*(4), 359–393. https://doi.org/10.1111/j.1468-0017.2010.01394.x

Sperber, D., & Wilson, D. (1995). *Relevance: Communication and Cognition*. Oxford: Blackwell Publishing. (Original work published 1986)

Sperber, D., & Wilson, D. (2006). Pragmatics. In E. Lepore & B. C. Smith (Eds.), *Oxford Handbook of Philosophy of Language.* Oxford University Press.

Starmans, C., & Friedman, O. (2012). The folk conception of knowledge. *Cognition*, *124*(3), 272–283. https://doi.org/10.1016/j.cognition.2012.05.017

Sternau, M., Ariel, M., & Giora, R. (2017). Deniability and explicatures. In R. Giora & M. Haugh (Eds.), *Doing Pragmatics Interculturally* (pp. 97–120). De Gruyter Mouton.

Sternau, M., Ariel, M., Giora, R., & Fein, O. (2015). Levels of interpretation: New tools for characterizing intended meanings. *Journal of Pragmatics*, *84*, 86–101. https://doi.org/10.1016/j.pragma.2015.05.002

Stokke, A. (2018). *Lying and Insincerity*. Oxford University Press. https://doi.org/10.1093/oso/9780198825968.001.0001

Stracqualursi, V., & Westwood, S. (2020, June 29). Trump thanked 'great people' shown in Twitter video in which a man chants 'white power'. *CNN*. https://edition.cnn.com/2020/06/28/politics/trump-tweet-supporters-man-chants-white-power/index.html

Sugden, R. (2000). The motivating power of expectations. In J. Nida-Rümelin & W. Spohn (Eds.), *Rationality, rules, and structure* (pp. 103–129). Dodrecht: Springer.

Szegőfi, Á., & Heintz, C. (2022). *Institutions of epistemic vigilance: The case of the newspaper press* [Manuscript in preparation]. https://philpapers.org/rec/SZEIOE-2

Székely, M., & Michael, J. (2018). Investing in commitment: Persistence in a joint action is enhanced by the perception of a partner's effort. *Cognition*, *174*(September 2017), 37–42. https://doi.org/10.1016/j.cognition.2018.01.012

Székely, M., Powell, H., Vannucci, F., Rea, F., Sciutti, A., & Michael, J. (2019). The perception of a robot partner's effort elicits a sense of commitment to human-robot interaction. *Interaction Studies*, *20*(2), 234–255.

The jamovi project. (2021). *Jamovi* (Version 1.6.23.0) [Computer software]. http://www.jamovi.org

Thomas, K. A., DeScioli, P., & Haque, O. S. (2014). The Psychology of Coordination and Common Knowledge. *Journal of Personality and Social Psychology*, *107*(4), 657–676. https://doi.org/10.1037/a0037037

Thomas, K. A., DeScioli, P., & Pinker, S. (2017). Common knowledge, coordination, and the logic of self-conscious emotions. *Evolution and Human Behavior*. https://doi.org/10.1016/j.evolhumbehav.2017.12.001

Tomasello, M. (1999). *The Cultural Origins of Human Cognition*. Harvard University Press. https://www.hup.harvard.edu/catalog.php?isbn=9780674005822

Tomasello, M. (2008). *Origins of human communication* (pp. xiii, 393). MIT Press.

Tomasello, M. (2009). *Why we cooperate?* Cambridge, MA: The MIT Press. https://doi.org/10.1002/hrm.20395

Tomasello, M. (2016). *A natural history of human morality*. Cambridge, MA: Harvard University Press.

Tomasello, M. (2018). The Normative Turn in Early Moral Development. *Human Development*, *61*(4–5), 248–263. https://doi.org/10.1159/000492802

Tomasello, M. (2020). The moral psychology of obligation. *Behavioral and Brain Sciences*, *43*, e56. https://doi.org/10.1017/S0140525X19001742.

Trump, D. J. [@realDonaldTrump]. (2017, May 30). *Despite the constant negative press covfefe* [Twitter]. https://web.archive.org/web/20170531054122/https://twitter.com/realDonaldTrump/status/869766994899468288

Tuomela, R. (2007). *The Philosophy of Sociality: The Shared Point of View*. Oup Usa.

Tuomela, R., & Miller, K. (1988). We-Intentions. *Philosophical Studies*, *53*(3), 367–389. https://doi.org/10.1007/BF00353512

Vaish, A., Missana, M., & Tomasello, M. (2011). Three-year-old children intervene in third-party moral transgressions: Children intervene in moral transgressions. *British Journal of Developmental Psychology*, *29*(1), 124–130. https://doi.org/10.1348/026151010X532888

Valdes, M. (2013, April 22). Innocents accused in online manhunt. *3 News*. https://web.archive.org/web/20131215111626/http://www.3news.co.nz/Innocents-accused-in-online-manhunt/tabid/412/articleID/295143/Default.aspx

Van Der Henst, J., Carles, L., & Sperber, D. (2002). Truthfulness and Relevance in Telling The Time. *Mind & Language*, *17*(5), 457–466. https://doi.org/10.1111/1468-0017.00207

Vanberg, C. (2008). Why Do People Keep Their Promises? An Experimental Test of Two Explanations. *Econometrica*, *76*(6), 1467–1480. https://doi.org/10.3982/ECTA7673

Vanderschraaf, P., & Sillari, G. (2014). Common Knowledge. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2014). http://plato.stanford.edu/archives/spr2014/entries/common-knowledge/

Viebahn, E. (2017). Non-literal Lies. *Erkenntnis*, *82*(6), 1367–1380. https://doi.org/10.1007/s10670-017-9880-8

Viebahn, E. (2019). Lying with Pictures. *The British Journal of Aesthetics*, *59*(3), 243–257. https://doi.org/10.1093/aesthj/ayz008

Viebahn, E., Wiegmann, A., & Willemsen, P. (2021). Can a question be a lie? An empirical investigation. *Ergo*, *8*(7).

Vignolo, A., Sciutti, A., Rea, F., & Michael, J. (2019). Spatiotemporal Coordination Supports a Sense of Commitment in Human-Robot Interaction. In M. A. Salichs, S. S. Ge, E. I. Barakova, J.-J. Cabibihan, A. R. Wagner, A. Castro-Gonzalez, & H. He (Eds.), *Social Robotics. ICSR 2019. Lecture Notes in Computer Science* (Vol. 11876, pp. 34–43). Springer International Publishing. https://doi.org/10.1007/978-3-030-35888-4_4

Vullioud, C., Clément, F., Scott-Phillips, T., & Mercier, H. (2017). Confidence as an expression of commitment: Why misplaced expressions of confidence backfire. *Evolution and Human Behavior*, *38*(1), 9–17. https://doi.org/10.1016/j.evolhumbehav.2016.06.002

Warneken, F., Gräfenhain, M., & Tomasello, M. (2012). Collaborative partner or social tool? New evidence for young children's understanding of joint intentions in collaborative activities. *Developmental Science*, *15*(1), 54–61. https://doi.org/10.1111/j.1467-7687.2011.01107.x

Warneken, F., & Tomasello, M. (2013). The emergence of contingent reciprocity in young children. *Journal of Experimental Child Psychology*, *116*, 338–350. https://doi.org/10.1016/j.jecp.2013.06.002

Weissman, B., & Terkourafi, M. (2019). Are false implicatures lies? An empirical investigation. *Mind & Language*, *34*(2), 221–246. https://doi.org/10.1111/mila.12212

Wiegmann, A., Samland, J., & Waldmann, M. R. (2016). Lying despite telling the truth. *Cognition*, *150*, 37–42. https://doi.org/10.1016/j.cognition.2016.01.017

Wiegmann, A., Willemsen, P., & Meibauer, J. (2021). *Lying, Deceptive Implicatures, and Commitment* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/n96eb

Willemsen, P., & Wiegmann, A. (2017). How the truth can make a great lie: An empirical investigation of the folk concept of lying by falsely implicating. *Proceedings of the 39th Annual Conference of the Cognitive Science Society*, 3516–3621. https://www.pascalewillemsen.com/wp-content/uploads/2017/12/Wiegmann-Willemsen_How-the-truth-can-make-a-great-lie.pdf

Wilson, D., & Sperber, D. (2002). Truthfulness and relevance. *Mind*, *111*(443), 583–632.

Wilson, D., & Sperber, D. (2004). Relevance Theory. In L. Horn & G. Ward (Eds.), *Handbook of Pragmatics* (pp. 607–632). John Wiley & Sons, Ltd.

Wilson, D., & Sperber, D. (2012). *Meaning and Relevance*. Cambridge University Press. https://doi.org/10.1017/CBO9781107415324.004

Wiltermuth, S. S., & Heath, C. (2009). Synchrony and Cooperation. *Psychological Science*, *20*(1), 1–5. https://doi.org/10.1111/j.1467-9280.2008.02253.x

Wyman, E., Rakoczy, H., & Tomasello, M. (2009). Normativity and context in young children's pretend play. *Cognitive Development*, *24*(2), 146–155. https://doi.org/10.1016/j.cogdev.2009.01.003

Xie, C., Yus, F., & Haberland, H. (Eds.). (2021). *Approaches to Internet Pragmatics*. John Benjamins Publishing Company.

Yamagishi, T., & Yamagishi, M. (1998). Trust and commitment as alternative responses to social uncertainty. *Networks, Markets, and the Pacific Rim: Studies in Strategy*, 109–124.

Yuan, W., & Lyu, S. (2022). Speech act matters: Commitment to what's said or what's implicated differs in the case of assertion and promise. *Journal of Pragmatics*, *191*, 128–142. https://doi.org/10.1016/j.pragma.2022.01.012

Zoom Video Communications. (2022). *Zoom* (5.10.1) [Computer software]. http://zoom.us

ACKNOWLEDGEMENT TO EXTERNAL FUNDING AGENCIES CONTRIBUTING TO PHD DISSERTATIONS

Name of doctoral candidate: *Francesca Bonalumi*

Title of dissertation: How we rely on each other: The perception of commitment in joint activities and communication

Name of supervisor(s): Christophe Heintz, Gergely Csibra
Name of advisor(s): John Michael, Thom Scott-Phillips

External funding agency: European Research Council

Acknowledgement: This research has received funding from the European Research Council (ERC) under the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement No 679092– SENSE OF COMMITMENT and Horizon 2020 programme (E/R/H20/2014-2020) under grant agreement No 742231 – PARTNERS.