

What makes an inference robust?*

Paul Marty
L-Università Ta' Malta
p.marty@ucl.ac.uk

Jacopo Romoli
Heinrich Heine Universität Düsseldorf
jacopo.romoli@hhu.de

Yasutada Sudo
University College London
y.sudo@ucl.ac.uk

Richard Breheny
University College London
r.breheny@ucl.ac.uk

Abstract Sentences involving embedded disjunctions give rise to DISTRIBUTIVE and FREE CHOICE inferences. These inferences exhibit certain characteristics of Scalar Implicatures (SIs) and some researchers have proposed to treat them as such. This proposal, however, faces an important challenge: experimental results have shown that the two inferences are more robust, faster to process, and easier to acquire than regular SIs. A common response to this challenge has been to hypothesise that such discrepancies among different types of SIs stems from the type of alternatives used to derive them. That is, in contrast to regular SIs, DISTRIBUTIVE and FREE CHOICE inferences are computed on the basis of sub-constituent alternatives, which are alternatives that are formed without lexical substitutions. This paper reports on a series of experiments that tested this hypothesis by comparing positive, disjunctive sentences giving rise to the two inference types to variants of these sentences involving either negation and conjunction, or negation and disjunction, for which the implicature approach predicts similar inferences on the basis of the same type of alternatives. The investigation also included deontic and epistemic modality, different positions of negation, and was extended to similar comparisons with simple disjunctions and the related IGNORANCE inferences they give rise to. Our results show that, while the inferences are indeed quite robust in the disjunctive cases, regardless of whether negation is present or not, the inferences that their negative, conjunctive variants give rise to are not. These findings are challenging for the hypothesis that the type of alternatives involved in SI computation is a major factor responsible for differences in robustness. We outline two possible alternative explanations of our data.

* For very helpful discussion and feedback, we would like to thank Maria Aloni, Moysh Bar-Lev, Kyle Blumberg, Marco Degano, Patrick Elliott, Danny Fox, Simon Goldstein, Matt Mandelkern, Sonia Ramotowska, Paolo Santorio, Uli Sauerland, Todd Snider and the audience at the ‘Scales, degrees, and implicature Workshop’ at Potsdam University, and at the Amsterdam Colloquium 2022. This research was supported by the Leverhulme Trust grant RPG-2018-425. Data files associated with our experiments, as well as the codes relevant to data treatment and statistical analyses that we report on are available open access on the OSF: <https://osf.io/x4v2a/>.

Keywords: disjunction, free choice, distributive inferences, ignorance inferences, negation, alternatives, relevance, alternative possibilities, implicature

Contents

1	Background	5
1.1	The implicature approach to Free choice and Distributive inferences	5
1.1.1	Basic ingredients	6
1.1.2	Distributive inferences	7
1.1.3	Free choice inferences	8
1.2	The challenge and the response	9
1.3	Novel predictions	12
2	Overview of the experiments	15
3	Experiments 1-3: Deontic modals	18
3.1	Participants	18
3.2	Materials	18
3.3	Procedure	21
3.4	Data treatment	23
3.5	Data analyses	23
3.6	Results	24
3.6.1	Control trials	24
3.6.2	Experiment 1: High Negation	24
3.6.3	Experiment 2: Intermediate Negation	27
3.6.4	Experiment 3: Low Negation	28
3.7	Discussion	28
4	Experiments 4-6: Epistemic modals and ignorance	29
4.1	Participants	31
4.2	Materials	32
4.3	Procedure	35
4.4	Data treatment	35
4.5	Data analyses	36
4.6	Results	36
4.6.1	Experiment 4: High negation	36
4.6.2	Experiment 5: Intermediate negation	39
4.6.3	Experiment 6: Low negation	40
4.7	Discussion	41
5	General discussion	44
6	Two directions	46
6.1	A relevance-based approach	46

6.1.1	Relevance and disjunction	46
6.1.2	Positive and low negative cases	48
6.1.3	High and intermediate negative cases	49
6.1.4	A note on regular SIs	49
6.1.5	Summary	50
6.2	A hybrid approach	50
6.2.1	The non-implicature part	51
6.2.2	Free choice inferences	52
6.2.3	Distributive inferences	53
6.2.4	Ignorance inferences	54
6.2.5	High and intermediate negation	55
6.2.6	Low negation	57
6.2.7	Summary	58
7	Conclusion	58
A	Training trials and Instructions in Exp.1-3	61
A.1	Instructions for Exp.1-3	61
A.2	Training trials in Exp.1-3	62
B	Training trials and Instructions in Exp.4-6	62
B.1	Instructions for Exp.4-6	62
B.2	Training trials in Exp.4-6	64

What makes an inference robust?

1 Background

Sentences like those in (1a) and (2a), where disjunction is embedded under a modal, give rise to the inferences in (1b) and (2b), respectively. The inference type in (1b) is generally referred to as a FREE CHOICE inference (henceforth, FC), and the one in (2b) as a DISTRIBUTIVE inference (henceforth, DI).¹

(1) FREE CHOICE (FC)

- a. It is possible that the box contains either a blue ball or a yellow ball.
- b. \rightsquigarrow *It's possible that the box contains a blue ball and it's possible that it contains a yellow ball*

(2) DISTRIBUTIVE INFERENCE (DI)

- a. It is certain that the box contains either a blue ball or a yellow ball.
- b. \rightsquigarrow *It's possible that the box contains a blue ball and it's possible that it contains a yellow ball*

These two inferences have been long known to display certain similarities to regular SCALAR IMPLICATURES (henceforth, SIs) like the one in (3b), arising from the simple sentence containing a possibility modal in (3a).

(3) SCALAR IMPLICATURE (SI)

- a. It is possible that the box contains a blue ball.
- b. \rightsquigarrow *It's not certain that the box contains a blue ball*

This fact has led some researchers to treat FC and DI as instances of SIs. In this section, we start by sketching this approach. Next, we discuss potential challenges coming from certain differences between these inferences and regular SIs arising from recent experimental research, and the main response to those challenges.

1.1 The implicature approach to Free choice and Distributive inferences

Implicatures were first extensively discussed by Grice (1975), who also coined the term *implicature* itself, and ever since then, the question of how implicatures arise has never been settled. A particularly well-discussed cases of implicatures are so-called SCALAR IMPLICATURES (SIs) like (3b) above. In a nutshell, the question is whether SIs arise mainly from the pragmatic side of the semantics-pragmatics interface by dint of pragmatic principles, or are computed in the semantic component of grammar as part of the conventional meaning of the linguistic expressions

¹ For FREE CHOICE inferences, see Kamp 1974, 1978 and much subsequent work; for DISTRIBUTIVE inferences, see Crnič, Chemla & Fox 2015, Chierchia 2013, Santorio & Romoli 2017, Ramotowska et al. 2022 among others.

used.² The debate between these two approaches is not crucial for our present purposes. In the following, we sketch a version of the latter approach, but we note that this choice is just for concreteness.

1.1.1 Basic ingredients

Let us follow Chierchia, Fox & Spector (2012) and much related work and assume that SIs arise from the application of a silent exhaustification operator, ‘EXH’, which combines with a sentence and returns the meaning of that sentence together with its SIs. More precisely, EXH takes as arguments a sentence and a set of contextually relevant alternatives to the sentence, and it returns the conjunction of that sentence with the negations of a subset of its relevant alternatives, namely those relevant alternatives that are ‘innocently excludable.’ In effect, EXH strengthens the meaning of the sentence, while avoiding contradictions and arbitrary choices between alternatives. The definition of EXH is given in (4) and that of innocent exclusion in (5), where ‘ C ’ stands for the set of salient alternatives.

$$(4) \quad \llbracket \text{EXH} \rrbracket(C)(p)(w) = p(w) \wedge \forall \phi \in \text{IE}(p, C) [\neg \llbracket \phi \rrbracket(w)]$$

$$(5) \quad \text{IE}(p, C) := \left\{ C' \mid \begin{array}{l} C' \text{ is a maximal subset of } C \text{ such that for some } w, \\ p \text{ is true at } w \text{ and each member of } C' \text{ is false at } w \end{array} \right\}$$

With these definitions in place, we can now illustrate how regular SIs are derived on this approach. For these purposes, consider the possible parse of (3) in (6), which involves EXH. The alternatives for EXH’s prejacent are given in (7).³ In this case, there is only one maximal excludable subset which includes only one alternative, namely the alternative labeled as ‘ $\Box a$ ’ below. Excluding this alternative gives rise to the intuitively correct implicature in (3b), namely *it is not certain that the box contains a blue ball*.

$$(6) \quad \text{EXH}[\text{It is possible that the box contains a blue ball}] = \Diamond a \wedge \boxed{\neg \Box a}$$

$$(7) \quad \left\{ \begin{array}{ll} \text{It is possible that the box contains a blue ball} & \Diamond a \\ \text{It is certain that the box contains a blue ball} & \Box a \end{array} \right\}$$

The meaning strengthening mechanism described here is very general and its application extends to a variety of simple and more complex cases involving regular SIs. In the following, we explain how it can be used to derive DI and FC.

² See, e.g., Sauerland 2004, Chierchia, Fox & Spector 2012, Franke 2011, Chemla 2010, Geurts 2010.

³ While the question of how alternatives are determined is still very debated, most theories of alternatives assume that the set of alternatives to (6) include those in (7); see Breheny et al. 2018 and references therein for discussion.

What makes an inference robust?

1.1.2 Distributive inferences

Consider again the case of DI in (2) above. The literal meaning of (2a) can be schematically represented as $\Box(a \vee b)$. This meaning does not entail the DISTRIBUTIVE inference and it is compatible with a situation in which one of the disjuncts is certain and the other impossible. However, if (2a) is parsed with EXH, as shown in (8), then the inference in (2b) arises as an implicature. To illustrate, assume that the prejacent of EXH, (2a), has the alternatives in (9).⁴

- (8) EXH[It is certain that the box contains a blue ball or a yellow ball]
- (9) $\left\{ \begin{array}{ll} \text{It is certain that the box contains a blue ball or a yellow ball} & \Box(a \vee b) \\ \text{It is certain that the box contains a blue ball} & \Box a \\ \text{It is certain that the box contains a yellow ball} & \Box b \\ \text{It is certain that the box contains a blue ball and a yellow ball} & \Box(a \wedge b) \end{array} \right\}$

All the alternatives in (9) but the prejacent are innocently excludable. The exclusion of these alternatives, and in particular the exclusion of the two alternatives corresponding to the modalised disjuncts, $\Box a$ and $\Box b$, yields the desired distributive meaning: it is possible that the box contains a blue ball and it is possible that it contains a yellow ball.

- (10) $\llbracket \text{EXH[It is certain that the box contains a blue ball or a yellow ball]} \rrbracket =$
 $\Box(a \vee b) \wedge \neg \Box(a \wedge b) \wedge \neg \Box a \wedge \neg \Box b =$
 $\Box(a \vee b) \wedge \neg \Box(a \wedge b) \wedge \boxed{\Diamond a \wedge \Diamond b}$

In sum, a theory of scalar implicatures can account for DISTRIBUTIVE inferences. As one can verify, the key observation here is that DISTRIBUTIVE inferences can be derived as SIs as long as each disjunct in the scope of the modal is taken to be an alternative. We turn next to the case of free choice.

⁴ The implicature approach to ‘distributive inference’ we are sketching here is the standard one, which derives them through the implication that each disjunct is not certain, as shown in (10). Recent work has unveiled cases where the distributive implication that each disjunct is possible arises (i.e., $\Diamond a \wedge \Diamond b$), without the implication that neither is certain (i.e., $\neg \Box a \wedge \neg \Box b$). This is challenging for the standard approach, because under this approach the two implications are necessarily linked i.e., one is derived from the other. As a result, alternative implicature accounts to distributive inferences, which can derive one implication without the other, have been put forward in the literature (Crnič, Chemla & Fox 2015, Bar-Lev & Fox 2020 among others). While we think this is a critical debate pushing our understanding of distributive inferences forward, here and in our experiments, we decided not to test situations that would pull the two implications apart. We therefore put this distinction aside for the time being; See Crnič, Chemla & Fox 2015, Bar-Lev & Fox 2020 and Ramotowska et al. 2022 for discussion.

1.1.3 Free choice inferences

The exhaustification-based process deriving regular SIs does not immediately derive FREE CHOICE inferences like those in (1), but it can be amended (and has been amended) in various ways to achieve this purpose (among others, Fox 2007, Klinedinst 2007, Santorio & Romoli 2017, Bar-Lev 2018, Bar-Lev & Fox 2020, Chemla 2010). Most prominently, Fox (2007) proposes that FC is a recursive or higher order implicature, arising through two successive applications of EXH. More recently, Bar-Lev (2018) and Bar-Lev & Fox (2020) put forward an amendment to the definition of EXH so that EXH not only excludes alternatives but also includes some others. We use this latter account for illustrative purposes but, here again, we note that nothing hinges on this presentation choice.

The proposal from Bar-Lev (2018) and Bar-Lev & Fox (2020) is that, in addition to conjoining the prejacent with the negation of its innocently excludable alternatives, the exhaustivity operator also conjoins the prejacent with a subset of other alternatives, those that are ‘innocently includable’. Concretely, innocently includable alternatives, as defined in (11), are those alternatives that are in all maximal subsets of alternatives that can be conjoined consistently with the assertion and with the negation of all innocently excludable alternatives. Following this characterisation, the definition of EXH is to be amended as shown in (12): EXH conjoins the prejacent with the negation of all innocently excludable alternatives, just as before, and now also conjoins it with all the innocently includable alternatives.

$$(11) \quad \Pi(p, C) := \left\{ C'' \left| \begin{array}{l} C'' \text{ is a maximal subset of } C \text{ such that for some } w, \\ p \text{ is true at } w \\ \text{and each member of } \text{IE}(p, C) \text{ is false at } w \\ \text{and each member of } C'' \text{ is true at } w \end{array} \right. \right\}$$

$$(12) \quad \llbracket \text{EXH} \rrbracket(C)(p)(w) = p(w) \wedge \forall \phi \in \text{IE}(p, C) [\neg \llbracket \phi \rrbracket(w)] \wedge \forall \phi \in \Pi(p, C) [\llbracket \phi \rrbracket(w)]$$

Assuming the novel definition in (12), FC inferences can be derived via a single application of EXH. To illustrate, consider the parse of (1a) in (13), where EXH occurs at matrix level. We assume that the alternatives to EXH’s prejacent are those in (14).

$$(13) \quad \text{EXH}[\text{It is possible that the box contains a blue ball or a yellow ball}]$$

$$(14) \quad \left\{ \begin{array}{ll} \text{It is possible that the box contains a blue ball or a yellow ball} & \diamond(a \vee b) \\ \text{It is possible that the box contains a blue ball} & \diamond a \\ \text{It is possible that the box contains a yellow ball} & \diamond b \\ \text{It is possible that the box contains a blue ball and a yellow ball} & \diamond(a \wedge b) \end{array} \right\}$$

The conjunctive alternative, namely $\diamond(a \wedge b)$, is the only innocently excludable alternative in this case. But there is now another way whereby the basic meaning of

What makes an inference robust?

the assertion can be strengthened, i.e., by identifying and including the innocently includable alternatives. As Bar-Lev & Fox (2020) show, there is one and only one subset of includable alternatives here, $\{\diamond(a \vee b), \diamond a, \diamond b\}$. Including each of these alternatives gives us the FC inference we were after, as shown in (15).

$$(15) \quad \llbracket \text{EXH}[\text{It is possible that the box contains a blue ball or a yellow ball}] \rrbracket = \diamond(a \vee b) \wedge \neg \diamond(a \wedge b) \wedge \boxed{\diamond a \wedge \diamond b}$$

Thus, the implicature approach can be extended to capture FC as well. Crucially, despite some differences between the derivations of DI and FC, it is easy to see that there is a striking similarity in the type of alternatives that these inferences are assumed to be derived from: just like DI, FC is derived on the basis of those alternatives that correspond to the modalised disjuncts (here, $\diamond a$ and $\diamond b$).

To summarise, the implicature approach has been a prominent approach in the literature and it can account for the two inferences in focus here: DISTRIBUTIVE and FREE CHOICE inferences. In fact, these inferences, and FREE CHOICE in particular, are commonly taken as a testing ground for disentangling predictions of different theories of implicatures (Fox 2007, Geurts 2010, Franke 2011, Bar-Lev 2018, Chemla 2010, Marty & Romoli 2021 among others). As we shall now see, however, results from recent experimental studies have challenged the implicature approach to DI and FC based on certain discrepancies between them and regular SIs.

1.2 The challenge and the response

Experimental work has unveiled discrepancies in the processing and acquisition of FC and DI, compared to regular SIs. In particular, FC has been consistently found to exhibit such differences. First, it has been shown to differ in its processing profile from regular SIs: Chemla & Bott 2014, building on Bott & Noveck 2004, found that responses based on FC interpretations are not slower than those based on the corresponding literal meanings, unlike what is usually found for regular SIs (see Van Tiel & Schaeken 2017 for similar results). Second, Tieu et al. (2016) found that 5-year-old children behave in a more adult-like fashion with FC than regular SIs, suggesting that the former type of inference is acquired earlier than the latter. DISTRIBUTIVE inferences have been studied less, but similar differences in processing speed and acquisition have been found for them as well (Van Tiel & Schaeken 2017, Pagliarini et al. 2018). That fact that FC and DI appear to be faster to process and easier to acquire than regular SIs raise a challenge for the implicature approach treating these inferences like regular SIs.

Thus far, the response in the literature has been to relate the observed discrepancies to the nature of the alternatives involved in the derivation of these inferences. The gist of the idea is that the derivation of regular SIs involves alter-

natives which requires lexical substitution, while those of DI and FC do not. To illustrate, consider again the example in (3) repeated from above. As it is easy to see, in order to arrive at the expected result in (3b), speakers need to entertain the alternative in (16), which involves substituting *possible* with its stronger scalemate, *certain*.

- (3) SCALAR IMPLICATURE (SI)
- a. It is **possible** that the box contains a blue ball.
 - b. \rightsquigarrow *It's not certain that the box contains a blue ball*
- (16) It is **certain** that the box contains a blue ball.

By contrast, none of the alternatives involved in the derivation of DI and FC requires lexical substitution. Rather, the derivations of these inferences are all based on alternatives involving constituents of the asserted sentence. For instance, as illustrated above, the DISTRIBUTIVE inferences associated with (2a), $\diamond a \wedge \diamond b$, are derived on the basis of the alternatives in blue in (17), corresponding to the disjuncts appearing in the scope of the modal.

- (2) DISTRIBUTIVE INFERENCE (DI)
- a. **It is certain that the box contains either a blue ball or a yellow ball.**
 - b. \rightsquigarrow *It's possible that the box contains a blue ball and it's possible that it contains a yellow ball*
- (17) $\left\{ \begin{array}{ll} \text{It is certain that the box contains a blue ball or a yellow ball} & \square(a \vee b) \\ \text{It is certain that the box contains a blue ball} & \square a \\ \text{It is certain that the box contains a yellow ball} & \square b \\ \text{It is certain that the box contains a blue ball and a yellow ball} & \square(a \wedge b) \end{array} \right\}$

Similar observations hold for the FREE CHOICE inferences associated with (1), $\diamond a \wedge \diamond b$, as evidenced below:

- (1) FREE CHOICE (FC)
- a. **It is possible that the box contains either a blue ball or a yellow ball.**
 - b. \rightsquigarrow *It's possible that the box contains a blue ball and it's possible that it contains a yellow ball*
- (18) $\left\{ \begin{array}{ll} \text{It is possible that the box contains a blue ball or a yellow ball} & \diamond(a \vee b) \\ \text{It is possible that the box contains a blue ball} & \diamond a \\ \text{It is possible that the box contains a yellow ball} & \diamond b \\ \text{It is possible that the box contains a blue ball and a yellow ball} & \diamond(a \wedge b) \end{array} \right\}$

What makes an inference robust?

In sum, the observation here is that the kind of alternatives involved in the computation of regular SIs differ from that involved in the derivation of DI and FC: the former involve lexical substitutions while the latter don't. Based on this observation, researchers have hypothesised that appealing to lexical substitution when building alternatives is what slows down the processing of an SI and makes it harder to acquire. This hypothesis can be formulated as follows (see Singh et al. 2016, Tieu et al. 2016, Pagliarini et al. 2018, Chemla & Bott 2014, Van Tiel & Schaeken 2017, Barner, Brooks & Bale 2011):

(19) **Alternative-based hypothesis (restricted)**

Alternatives that do not involve lexical substitutions give rise to inferences that are faster to process and easier to acquire.

The Alternative-based hypothesis above has been primarily formulated in reference to processing and acquisition results. Yet another way in which DI and FC have been shown to differ from regular SIs is in their robustness (Chemla 2009, Van Tiel & Schaeken 2017, Marty et al. 2021). Thus, it is natural to extend the original Alternative-based hypothesis along the lines of (20) so as to include robustness as well.

(20) **Alternative-based hypothesis (extended)**

Alternatives that do not involve lexical substitutions give rise to inferences that are more robust, faster to process and easier to acquire.

This version of the Alternative-based hypothesis permits to reconcile the implicature approach with the experimental results pertaining to the robustness of DI and FC reported in the literature. As we shall see, this hypothesis also makes novel predictions. Before we turn to these predictions, let us emphasise that, while we formulated two separate hypotheses above, they both directly relate the discrepancies under discussion to a single source having to do with the nature of the alternatives involved in the derivation of the inferences in question. It also bears pointing out that, although the cognitive cost of a scalar inference is in principle independent from its robustness, it is reasonable to assume that the two measures are positively correlated to some extent. It could be so, for instance, if both of the following are the case: (i) The lower the processing cost associated with a given inference, the more likely it is that this inference will be accessed by speakers, and (ii) when speakers perceive both the strong and weak reading of a given sentence, they preferentially resolve the ambiguity at hand by favoring the stronger over the weaker reading. These assumptions link the likelihood that a given inference be accessed by speakers, which partly depends on its processing cost (among other

factors), to the likelihood that its corresponding reading be reported by speakers, which depends in turn on its accessibility. We come back to this point below.

1.3 Novel predictions

Once supplemented with the hypothesis in (20), the implicature approach predicts that inferences based on non-lexical alternatives should behave more like FC and DI, and unlike regular SIs. One straightforward way to test this general prediction is to look at the negative counterparts of (1a)–(2a) in (21)–(22) and their predicted inferences.

- (21) NEGATIVE FREE CHOICE
- a. It is not certain that the box contains both a blue ball and a yellow ball.
 - b. \sim *It's not certain that the box contains a blue ball and it's not certain that it contains a yellow ball*
- (22) NEGATIVE DISTRIBUTIVE INFERENCE
- a. It is not possible that the box contains both a blue ball and a yellow ball.
 - b. \sim *It's not certain that the box contains a blue ball and it's not certain that it contains a yellow ball*

These cases are particularly interesting for our purposes because the implicature approach derives negative FC and negative DI in the same way as their positive counterparts, on the basis of the same type of non-lexical alternatives. To illustrate, consider first the case of negative FC in (21). The literal meaning of this sentence, as predicted by standard approaches to modals and conjunction, is equivalent to $\neg\Box a \vee \neg\Box b$ and it is thus compatible with one of a or b being certain. However, in the same way as before, this literal meaning can be exhaustified as shown in (23), using the alternatives in (24) (where the critical alternatives are indicated in blue, as before).

- (23) EXH[It is not certain that the box contains both a blue ball and a yellow ball]
- (24) $\left\{ \begin{array}{l} \text{It is not certain that the box contains both a blue ball and a yellow ball} \\ \text{It is not certain that the box contains a blue ball} \\ \text{It is not certain that the box contains a yellow ball} \\ \text{It is not certain that the box contains either a blue ball or a yellow ball} \end{array} \right\}$ $\left. \begin{array}{l} \neg\Box(a \wedge b) \\ \neg\Box a \\ \neg\Box b \\ \neg\Box(a \vee b) \end{array} \right\}$

What makes an inference robust?

In a similar way as for positive FC, $\neg\Box(a \vee b)$ is the only innocently excludable alternative while the other three are innocently includable, i.e., $\Pi = \{\neg\Box(a \wedge b), \neg\Box a, \neg\Box b\}$. By including these alternatives, we obtain the negative free choice meaning of the sentence:

$$(25) \quad \llbracket \text{EXH}[\text{It's not certain that the box contains a blue and a yellow ball}] \rrbracket = \neg\Box(a \wedge b) \wedge \Box(a \vee b) \wedge \boxed{\neg\Box a \wedge \neg\Box b}$$

Importantly, note that the alternatives over which negative FC is derived are parts of the asserted sentence and do not involve any lexical substitution. The same can be shown for the negative DISTRIBUTIVE inference in (22), which is derived from the alternatives in (26). In particular, the alternatives in blue end up being excluded. Their exclusion, together with the literal meaning of the sentence, entails the negative DI inference indicated above.

$$(26) \quad \left\{ \begin{array}{l} \text{It is not possible that the box contains both a blue ball and a yellow ball} \\ \text{It is not possible that the box contains a blue ball} \\ \text{It is not possible that the box contains a yellow ball} \\ \text{It is not possible that the box contains either a blue ball or a yellow ball} \end{array} \right. \left. \begin{array}{l} \neg\Diamond(a \wedge b) \\ \neg\Diamond a \\ \neg\Diamond b \\ \neg\Diamond(a \vee b) \end{array} \right\}$$

$$(27) \quad \llbracket \text{EXH}[\text{It is not possible that the box contains a blue and a yellow ball}] \rrbracket = \neg\Diamond(a \wedge b) \wedge \Diamond(a \vee b) \wedge \Diamond a \wedge \Diamond b = \neg\Diamond(a \wedge b) \wedge \Diamond(a \vee b) \wedge \boxed{\neg\Box a \wedge \neg\Box b}$$

Finally, call the cases in (21) and (22) ‘high negative’ cases, and consider the equivalent ‘intermediate’ negative cases in (28a) and (29a), where negation appears just below the modal, and the corresponding ‘low’ negative versions in (30a) and (31a), where negation is embedded further down. Since the implicature approach works on logical relations between the asserted sentence and its alternatives, it predicts the availability of the same inferences for all these logically equivalent cases.

- (28) INTERMEDIATE NEGATIVE FC
- a. It is possible that the box does not contain both a blue ball and a yellow ball.
 - b. \rightsquigarrow *It's not certain that the box contains a blue ball and it's not certain that it contains a yellow ball*

- (29) INTERMEDIATE NEGATIVE DI
- a. It is certain that the box does not contain both a blue ball and a yellow ball.
 - b. \rightsquigarrow *It's not certain that the box contains a blue ball and it's not certain that it contains a yellow ball*

- (30) LOW NEGATIVE FC
- a. It is possible that the box either does not contain a blue ball or it does not contain a yellow ball.
 - b. \rightsquigarrow It's not certain that the box contains a blue ball and it's not certain that it contains a yellow ball
- (31) LOW NEGATIVE DI
- a. It is certain that the box either does not contain a blue ball or it does not contain a yellow ball.
 - b. \rightsquigarrow It's not certain that the box contains a blue ball and it's not certain that it contains a yellow ball

Table 1 summarises all the cases discussed so far, together with their predicted inferences, across the different sentence types and environments.

Environment	Schematic description		
	FC	DI	SI
POSITIVE	$\diamond(a \vee b)$	$\square(a \vee b)$	$\diamond a$
<i>Inference:</i>	$\diamond a \wedge \diamond b$	$\diamond a \wedge \diamond b$	$\neg \square a$
NEGATIVE			
High	$\neg \square(a \wedge b)$	$\neg \diamond(a \wedge b)$	$\neg \square a$
Intermediate	$\diamond \neg(a \wedge b)$	$\square \neg(a \wedge b)$	
Low	$\diamond(\neg a \vee \neg b)$	$\square(\neg a \vee \neg b)$	$\diamond \neg a$
<i>Inference:</i>	$\neg \square a \wedge \neg \square b$	$\neg \square a \wedge \neg \square b$	$\diamond a$

Table 1 Overview of the cases investigated in this study, along with their predicted inferences. The positive and the low negative cases involve disjunction, while the other cases involve conjunction.

As discussed, the implicature approach predicts the negative inferences described in this section to be derived in the same way and on the basis of the same type of alternatives as their positive counterparts. Coupled with the Alternative-based hypothesis, this approach predicts these inferences to behave like their positive counterparts, and unlike regular SIs, in terms of strength, processing, and acquisition. Moreover, these predictions do not change across the different variants of the negative cases that we have described, i.e., whether negation appears high in the sentence, at an intermediate position below the modal, or within each disjunct. These predictions are summarised in Table 2.

Environment	Inference type		
	FC	DI	SI
POSITIVE	✔	✔	✔
NEGATIVE			
High	✔	✔	✔
Intermediate	✔	✔	
Low	✔	✔	✔

Table 2 Predictions of the implicature approach, supplemented with the Alternative-based hypothesis in (20). A checkmark indicates that the relevant inference is predicted to be available. A dark green checkmark ✔ indicates that a robust inference is predicted, while a light green one ✔ stands for a weaker/regular inference.

In the following sections, we present and report on a series of experiments testing the above predictions for FC and DI. Before moving to the experiments, however, one last clarification is in order. Although the implicature approach predicts no difference between positive and negative cases, it is possible that negative sentences give rise to less robust implicatures, because these sentences are generally harder to process than their positive counterparts. To control for this potential effect of negation, it is thus critical to also take regular SIs into consideration, both in their positive and negative versions, and add these cases to our set of comparison points for FC and DI. If, indeed, negation lowers implicature derivation across the board, this effect should be observed for regular SIs as well.

2 Overview of the experiments

Some of the cases that we were interested in have been investigated in two previous studies. The first one is Chemla 2009, which compared positive and high negative FC in simple and quantificational environments. The second is Marty et al. 2021, which tested positive and high negative FC against various baselines and compared them to positive and negative SIs. Both of these studies found a clear difference in inference strength between positive FC and high negative FC cases involving deontic modalities (e.g., *allowed/required*). The present study aimed to expand on these initial investigations in two main ways.

First, we extended the empirical scope of previous studies by looking at the effect of sentence polarity across three inferences types – FC, DI and SI – and by

testing not only deontic modals, but also epistemic ones. Second, we experimentally manipulated the position of negation in the negative cases so as to cover the whole range of linguistic environments presented in Table 1 – high, intermediate and low negative cases. By adding these novel cases to the set of comparison points, our goal was to investigate the effect of sentence polarity on inference strength in a more systematic fashion, to reach a more complete and accurate description of this effect so as to assess the generality of this phenomenon, to evaluate the challenge it constitutes for the Alternative-based hypothesis and, to test the predictions of non-implicature approaches to FC and DI as well.

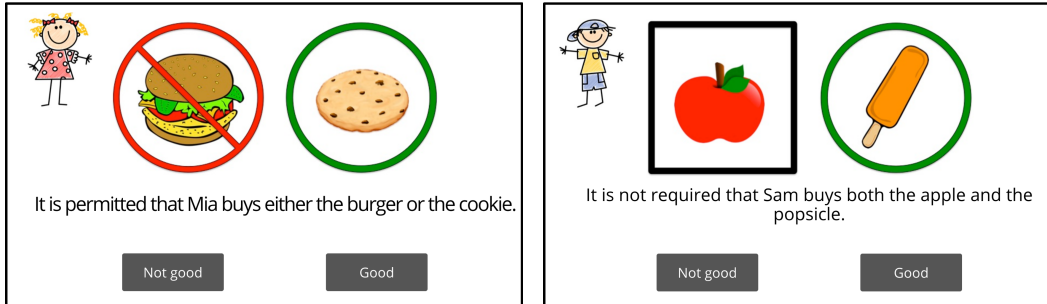
In the following two sections, we report on two sets of three experiments that we carried out to achieve these purposes. All our experiments involved a sentence-picture acceptability task where participants were presented with sentence-picture items like those in Figure 1, and had to decide whether the sentence was a good description of the situation depicted in the picture. Participants reported their judgement by clicking on one of two response buttons, labelled ‘Good’ and ‘Bad’ (or ‘Not good’), respectively. In the critical conditions, the test sentences were paired with pictures that make them false if the target inference is derived, but true if not, as illustrated in Figure 1 for positive and high negative FC. The linking hypothesis was that the rates of rejection (‘Bad/Not Good’ responses) observed in these conditions is a proxy for the robustness of the target inference: the more robust a given inference is, the more participants should select the ‘Bad’/‘Not good’ response option in these conditions and, consequently, the higher the rejection rate should be.

The first set of experiments (Exp.1-3, Section A) tested FC and DI involving deontic modalities (i.e., *permitted/required*), using the same paradigm as in Marty et al. (2021), while the second set of experiments (Exp.4-6, Section B) extended this investigation to FC and DI cases involving epistemic modalities (i.e., *possible/certain*), building on the covered box paradigm from Noveck (2001). For the sake of comparison, all three experiments in each set were designed in a similar fashion and differed from one another along one critical dimension, the level of the structure at which negation occurs in the negative sentences. In effect, the negative FC and DI sentences tested in each set of experiments were obtained from their positive counterparts by placing a negation at the matrix level (high), at the level of the main verb of the embedded clause (intermediate) or within each sub-clause of the embedded clause (low), as well as by replacing, whenever relevant, the embedded modals and/or connectives with their scalemates.

Exp.1-3 tested the predictions from the Alternative-based hypothesis (see Table 1) by investigating positive instances of FC and DI and their three negative variants and by comparing them to positive SI and its corresponding negative

What makes an inference robust?

Experiments 1-3: Deontic modals



Experiments 4-6: Epistemic modals

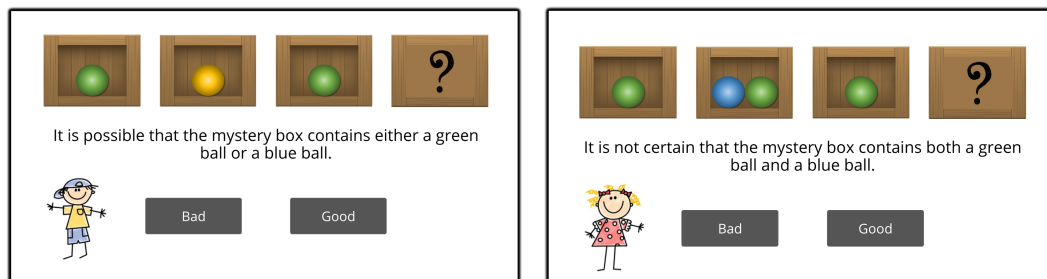


Figure 1 Examples of items used in Exp.1-3 (top) and Exp.4-6 (bottom). These examples correspond to target trials for the positive and negative FC sentences in Exp.1 and Exp.4, respectively. In Exp.1-3, participants were instructed to interpret the green circle as *allowed but not required*, the red circle as *not allowed*, and the black square as *required* (see Section 3.3 and Appendix A.1). In Exp.4-6, participants were instructed that the rightmost box, called *the mystery box*, always had the same contents as one of the three open boxes (see Section 4.3 and Appendix B.1).

variants. Exp.4-6 were parallel to Exp.1-3 and were primarily designed to test the epistemic equivalents of the positive and negative deontic sentences from Exp.1-3.

Finally, in Exp.4-6 we also added simple disjunctive sentences, together with their negative variants, as a way to test the strength of the so-called *IGNORANCE* inferences that these sentences give rise to. As we discuss in Section 4, these inferences have also been claimed to arise as *SIs* on the basis of sub-constituents alternatives, just like *FC* and *DI*. Therefore, we would expect *IGNORANCE* inferences to pattern with *FC* and *DI* in terms of robustness, as one would expect if (i) these inferences are regular *SIs* and (ii) the *Alternative-based hypothesis* is correct.

3 Experiments 1-3: Deontic modals

3.1 Participants

For each experiment, 60 participants were recruited online through Prolific using the same pre-screening criteria across the board (country of residence: UK; country of birth: UK; first language: English; minimum prior approval rate: 90%; vision: see colours normally). The recruitment was set up so that each participant could only take part in one of the three experiments. They were paid £1.75 for their participation and average completion time was around 12 minutes. All participants gave written informed consent to the processing of their information for the purposes of this study, which was approved by the institution's Research Ethics Committee. Data was collected and stored in accordance with the provisions of Data Protection Act 2018, the UK's implementation of the General Data Protection Regulation.

3.2 Materials

All three experiments were designed in a similar way based on the materials and method from Marty et al. (2021: Exp.1-2). Every picture displayed two different food items from the following list: *popsicle, pear, broccoli, cookie, pineapple, croissant, burger, hot-dog, cupcake, fries, pizza, donut, banana, carrot, apple, avocado, orange, lettuce*. Each food item was enclosed within one of three symbols: a green circle, a red circle with a red diagonal line through it, and a black square (see examples in Figure 1). Participants were instructed to interpret these symbols as representing specific rules concerning what two children, Mia and Sam, are *allowed but not required to buy* (green circle), *not allowed to buy* (red circle) and *required to buy* (black square) at the supermarket, according to their parents (see Appendix A.1). In each experiment, the 18 food items from the above list were evenly distributed among both characters and evenly combined with the three rule-related symbols by pseudo-randomly assigning to each character 3 food items that he was permitted to buy, 3 that he was not permitted to buy and 3 that he was required to buy. As a result, each food item was consistently associated with the same rule within the same experiment. Pictures were generated using the picture frames in Table 3 by randomly selecting two food items with the appropriate symbols and randomly determining their ordering on the picture.

Sentences were generated on the basis of the picture they were paired with using the sentence frames in Table 4. The [name] term was the name of one of the two characters, *Mia* or *Sam*. The [L] and [R] terms were food names corresponding to the names of the food items displayed on the *left* and on the *right* of the picture, respectively. Hence, the linear order of the food terms in the sentence always


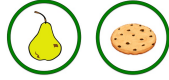


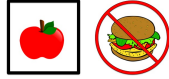
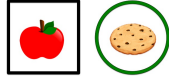
Label	Description of the picture type	Example picture
DD	not permitted to buy either items	
AA	permitted to buy both items	
RR	required to buy both items	
AD	permitted to buy one item & not permitted to buy the other	
RD	required to buy one item & not permitted to buy the other	
RA	permitted to buy one item & required to buy the other	

Table 3 Description and examples of the picture types used in Exp.1-3.

matched their (left-to-right) ordering on the picture. For sentence types involving only one food term (i.e., positive and negative SI), the food name used in the sentence was pseudo-randomly assigned to match one of the two food items displayed on the picture (i.e., the one on the left or the one on the right) in accordance with the experimental condition.

Exp.1 tested three positive sentences, one for each inference type of interest (FC, DI and SI), together with their high negative variants, obtained from the positive sentences by adding negation at matrix level and by replacing the embedded modals and connectives with their scalemates (i.e., *permitted*→*required*, *required*→*permitted*, *or*→*and*). Exp.2 tested the same positive sentences as in Exp.1 and compared them to their intermediate negative variants. These variants were obtained from the positive sentences by placing a negation at the level of the main verb of the embedded clause and by replacing, wherever relevant, the embedded connectives with their scalemate (i.e., *or*→*and*). Finally, Exp.3 tested the same positive sentences as in Exp.1-2 and compared them to their low negative variants, where negation is embedded as low as possible. The low negative variants of FC and DI were built after their positive forms by placing a negation in each sub-clause of the complex embedded clause. The contents of these sentences were further adjusted by making all the syntactic subjects phonologically explicit so as to ensure

Exp.1 – High negation

POSITIVE

- FC It is permitted that [name] buys either the [L] or the [R].
 DI It is required that [name] buys either the [L] or the [R].
 SI It is permitted that [name] buys the [L/R].

NEGATIVE

- FC It is not required that [name] buys both the [L] and the [R].
 DI It not permitted that [name] buys both the [L] and the [R].
 SI It is not required that [name] buys the [L/R].
-

Exp.2 – Intermediate negation

POSITIVE

- FC It is permitted that [name] buys either the [L] or the [R].
 DI It is required that [name] buys either the [L] or the [R].
 SI It is permitted that [name] buys the [L/R].

NEGATIVE

- FC It is permitted that [name] does not buy both the [L] and the [R].
 DI It is required that [name] does not buy both the [L] and the [R].
 SI It is permitted that [name] does not buy the [L/R].
-

Exp.3 – Low negation

POSITIVE

- FC It is permitted that [name] buys either the [L] or the [R].
 DI It is required that [name] buys either the [L] or the [R].
 SI It is permitted that [name] buys the [L/R].

NEGATIVE

- FC It is permitted that either [name] does not buy the [L]
 or [name] does not buy the [R].
 DI It is required that either [name] does not buy the [L]
 or [name] does not buy the [R].
 SI It is permitted that [name] does not buy the [L/R].
-

Table 4 Schematic description of the sentences tested Exp.1-3 by sentence polarity and inference type, where [name] stands for the name of one of the two characters and [L] and [R] stand for the names of the food items displayed respectively on the left and on the right of the pictures the sentences were paired with; for a more concrete illustration, you may read [name] as *Mia*, [L] as *burger* and [R] as *pizza*.

What makes an inference robust?

that they sound as natural as possible to native speakers. The negative SI sentences in Exp.3 were the same as in Exp.2.

The sentence-picture combinations were the same in all three experiments. These combinations are summarized and illustrated in Table 5. In the **Target** conditions, FC, DI and SI sentences were paired with pictures that make them false if the relevant inference is present, but true if it is absent. For the FC and SI sentences, control **True** and **False** conditions were further included by pairing these sentence with pictures that make them either unambiguously true or unambiguously false, regardless of the presence or absence of the inferences of interest. The inclusion of analogous conditions for the DI sentences was either not possible or problematic due to other considerations.⁵ Thus, FC and SI sentences were paired with three picture types, corresponding to their **True**, **False** and **Target** conditions, and DI sentences with only one picture type, corresponding to their **Target** conditions. For each sentence type, the **True** and **False** conditions (where available) were instantiated 4 times, and the **TARGET** conditions 10 times, resulting in 92 trials in each experiment.

3.3 Procedure

The three experiments were run as online surveys using Gorilla Experiment Builder (Anwyl-Irvine et al. 2020). The procedure and instructions were the same in all three experiments and the same as in Marty et al. (2021: Exp.1-2). In the instructions, participants were introduced to two characters, Sam and Mia, and they were presented with a short cover story (see Appendix A.1 for details). They were told that they would see pictures representing by means of different symbols the parental rules concerning what Sam and Mia are allowed to buy, not allowed to buy and required to buy at the supermarket. To illustrate the meaning of each symbol, participants were then shown three examples of enclosed food items, one for each symbol, together with their intended interpretation. Finally, participants were told that each picture in the study will be accompanied with a sentence that relates to it and they were instructed to click on ‘Good’ if they consider that sentence a good description of the picture they see and otherwise to click on ‘Not good’.

Following the instructions, participants started the survey with a short training devised to consolidate their understanding of the cover story and instructions (see

⁵ For the DI sentences, none of the picture types in Table 3 could be used to create analogous **True** conditions as none of them makes the positive or negative DI sentences *unambiguously true*. In particular, the RR pictures make the positive DI sentences false if they are understood with their exclusive SI and the DD pictures make their negative variants false if they are understood with the corresponding inclusive SI. While there was no particular issue for creating **False** conditions for these sentences, including only those conditions may have caused other potential problems, e.g., by making the design of our experiments slightly unbalanced.





























	True	False	Target
POSITIVE			
FC	  A A	  D D	  A D
DI	N/A		  R A
SI	  A D	  D D	  R D
NEGATIVE			
FC	  A A	  R R	  R A
DI	N/A		  A D
SI	  R A	  R R	  R D

Table 5 Summary of the sentence-picture combinations giving rise to the **True**, **False** and **Target** conditions in Exp.1-3. The position of the food items as well as of the symbols represented on the pictures was randomized. In these examples, the food name appearing in the SI sentences would correspond to the food item on the left for the positive instance and to the food item on the right for its negative counterpart.

Appendix A.2 for details). After the training phase, the survey continued with a block of 92 test trials. Trials were presented in random order, with a 1000 ms interstimulus interval. Participants reported their judgments by clicking with the mouse one of two response buttons labelled ‘Good’ and ‘Not Good’, respectively.

What makes an inference robust?

The position of the response buttons on the screen (i.e., on the left or on the right) was counterbalanced across participants. Items remained on the screen until participants gave their response. At the end of the survey, participants were asked to fill out a short demographic questionnaire.

3.4 Data treatment

Responses from participants whose performance in the **True** and **False** control trials did not reach the pre-established threshold of 70% accuracy were excluded prior to analyses. Responses from 2 participants in Exp.1, 7 participants in Exp.2 and 8 participants in Exp.3 were removed through this procedure, representing about 3%, 11% and 13% of the data collected in these experiments, respectively.

3.5 Data analyses

Data from each experiment was analysed through model comparison of binomial linear mixed-effects models (Jaeger 2008), by carrying out two main analyses. The first one focused on FC and aimed to assess the generality of the findings from Marty et al. 2021 by testing whether the positive instances of FC are more robust than their negative variants, as previously observed for high negative FC (Marty et al. 2021: Exp.1-2). The second focused on participant's responses in the **Target** conditions and aimed to generalize the first analysis by comparing the robustness of FC, DI and SI on their positive and negative instances.

Analyses were conducted using the lme4 (Bates et al. 2015), car (Fox & Weisberg 2019) and emmeans (Lenth 2021) libraries for the R statistics program (R Core Team 2021). In the first analysis, we assessed whether participants' responses in the **True** and **Target** conditions for FC differ as a function of sentence polarity. The models included Condition (2 levels: True, Target; sum-coded), Polarity (2 levels: Positive, Negative; sum-coded) and their interaction as fixed effects, with Subject as a random effect and a by-Subject random slope for both fixed effects and their interaction. In the second analysis, we assessed whether participants' responses in the **Target** conditions differ as a function of the inference type and sentence polarity. The models included Inference type (3 levels: SI, FC, DI; dummy-coded), Polarity (2 levels: Positive, Negative; sum-coded) and their interaction as fixed effects, with Subject as a random effect and a by-Subject random slope for both fixed effects and their interaction. Inference type was coded with treatment contrasts using SI as a fixed reference level, thus allowing us to use the regression output to compare FC and DI to SI. To complete these analyses, *post-hoc* pairwise comparisons were performed based on the estimated marginal means from the

models we just described; p -values were adjusted using the Bonferroni correction method for multiple testing.

3.6 Results

The bar plot in Figure 2 shows the mean rates of rejection for each sentence type, in each experiment, by condition, inference type and sentence polarity. In the **Target** conditions, the mean rejection rates are proxy measures for the robustness of the target inference: the higher the rejection rate, the more robust the corresponding inference. Zooming in, the interaction plot in Figure 3 shows the behavior of the mean rejection rates in the **Target** conditions for each sentence type as a function of sentence polarity.

3.6.1 Control trials

Responses to the control trials were as expected in all three experiments: participants uniformly rejected the FC and SI sentences in their **False** conditions (all mean rejection rates $\geq 95\%$) and strongly accepted them in their **True** conditions (all mean rejection rates $\leq 22\%$), with an overall mean accuracy of 94% (95% CI [93, 95]) in Exp.1, 92% (95% CI [93, 95]) in Exp.2 and 93% (95% CI [91, 94]) in Exp.3. These results show that participants did the task appropriately and, in particular, that they had no difficulty mapping each basic symbol to its intended meaning or interpreting their various combinations across trials.

3.6.2 Experiment 1: High Negation

The patterns of results for the positive and high negative instances of FC were similar to those reported in Marty et al. (2021: Exp.2). Specifically, the mean rejection rate for positive FC in the **Target** conditions was very high (M=95%, 95% CI [93, 96]), much higher than in the **True** conditions (M=6%, 95% CI [3, 9]) and as high as in the **False** conditions (M=97%, 95% CI [95, 99]). By contrast, in these same conditions, the mean rejection rate for negative FC was somewhat intermediate (M=40%, 95% CI [36, 44]), in between those observed in the corresponding control conditions (**True**: M=16%, 95% CI [12, 22]; **False**: M=97%, 95% CI [95, 99]). The model for FC yielded a significant effect for Condition ($\chi^2(1) = 38.37, p < .001$), Polarity ($\chi^2(1) = 10.55, p < .01$) and their interaction ($\chi^2(1) = 22.39, p < .001$) such that the difference between **True** and **Target** conditions was significantly larger for positive FC than for its high negative variant. These results reproduce those from Marty et al. (2021: Exp.2) in showing that FC inferences are available in

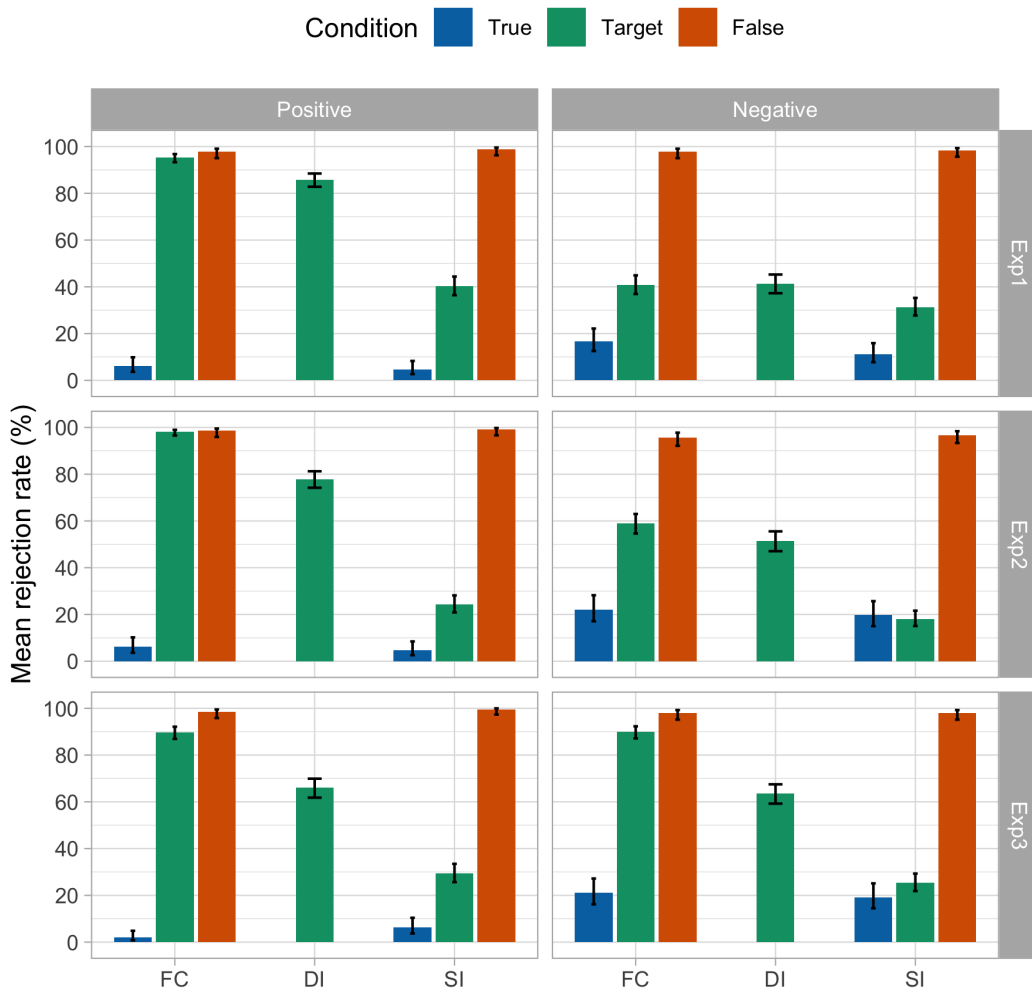


Figure 2 Mean rejection rate (in %) to each sentence type in Exp.1 (top), Exp.2 (middle) and Exp.3 (bottom) by sentence polarity and experimental condition. Error bars represent 95% confidence intervals.

their high negative form, but considerably less so than in their classical, positive form.

Turning now to the comparisons between inference types, DI was found to pattern with FC and distinctly from SI. In the **Target** conditions, positive DI gave rise to high rejection rates ($M=85\%$, 95% CI [82, 88]) and its high negative variant to intermediate ones ($M=41\%$, 95% CI [37, 45]), just like FC; by contrast, SI gave rise to intermediate rejection rates across-the-board (positive: $M=40\%$, 95% CI [36, 44]; negative: $M=31\%$, 95% CI [27, 35]). The model for the **Target** conditions yielded

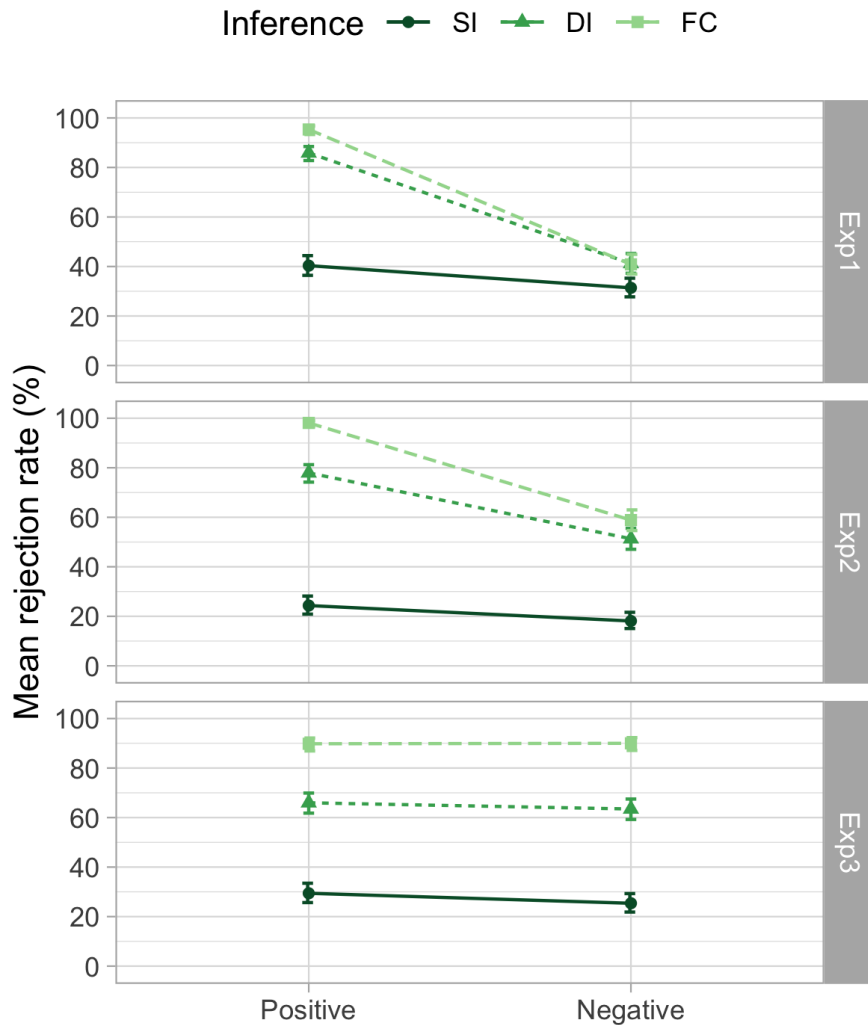


Figure 3 Mean rejection rates (in %) in the **Target** conditions of Exp.1 (top), Exp.2 (middle) and Exp.3 (bottom) by inference type and sentence polarity. Error bars represent 95% confidence intervals.

a significant effect of Inference type ($\chi^2(1) = 35.99, p < .001$) and a significant interaction between Inference type and Polarity ($\chi^2(1) = 25.68, p < .001$) such that the difference between positive and negative forms was larger for both FC ($\beta = 3.45, SE = 0.83, p < .001$) and DI ($\beta = 1.79, SE = 0.43, p < .001$) than for SI. The data was investigated further by comparing, for each inference type, the estimated marginal means for their positive and negative instances. The results

What makes an inference robust?

showed a significant contrast between positive and negative cases for both FC ($\beta = 7.62, SE = 1.60, p < .001$) and DI ($\beta = 4.29, SE = 0.74, p < .001$), but not for SI ($\beta = 0.71, SE = 0.57, p = 0.21$). Taken together, these findings establish that FC and DI were far more accessible than regular SIs in their positive forms and that, compared to their positive forms, both these inferences became far less accessible in their high negative forms, unlike regular SIs.

3.6.3 Experiment 2: Intermediate Negation

The results from Exp.2 were similar to those from Exp.1 across all factors of interest. Starting with the FC results, the mean rejection rate for positive FC in the **Target** conditions was very high (M=98%, 95% CI [96, 99]) whereas the mean rejection rate for negative FC in these same conditions was intermediate (M=58%, 95% CI [54, 62]). The model for FC yielded a significant effect for Condition ($\chi^2(1) = 53.59, p < .001$) and for the interaction between Condition and Polarity ($\chi^2(1) = 17.31, p < .001$) such that the difference between **True** and **Target** conditions was significantly larger for positive FC than for its intermediate negative variant. These results establish that FC inferences are also available in their intermediate negative form, but considerably less so than in their positive form, similar to what we found for the high negative cases.

As in Exp.1, DI was found to pattern more closely with FC than with SI. In the **Target** conditions, positive DI gave rise to high rejection rates (M=78%, 95% CI [74, 81]) whereas its high negative variant gave rise to intermediate ones (M=51%, 95% CI [47, 55]). In these same conditions, SI gave rise to relatively low rejection rates across-the-board (positive: M=24%, 95% CI [20, 28]; negative: M=18%, 95% CI [15, 21]). The model for the **Target** conditions yielded a significant effect of Inference type ($\chi^2(1) = 50.59, p < .001$) and a significant interaction between Inference type and Polarity ($\chi^2(1) = 6.50, p < .05$) such that the difference between positive and negative forms was larger for FC than SI ($\beta = 2.41, SE = 0.98, p < .05$); the comparison involving DI and SI, however, did not reach significance ($\beta = 0.93, SE = 0.61, p = .12$). The post-hoc pairwise comparisons yielded the same results as in Experiment 1: there was a significant contrast between positive and negative cases for both FC ($\beta = 6.20, SE = 1.78, p < .001$) and DI ($\beta = 3.24, SE = 0.87, p < .001$), but not for SI ($\beta = 1.37, SE = 0.89, p = 0.12$). In sum, the results from Exp.2 extend the main findings from Exp.1 to the novel comparisons between positive and intermediate negative cases.

3.6.4 Experiment 3: Low Negation

The results from Exp.3 stood in stark contrast with those from Exp.1-2: for all three inference types, the mean rejection rates for the low negative forms were similar to those for the positive forms. In the **Target** conditions, the mean rejection rates for FC were very high (positive: $M=89\%$, 95% CI [86, 92]; negative: $M=90\%$, 95% CI [87, 92]), those for DI were in the upper mid-range (positive: $M=65\%$, 95% CI [61, 69]; negative: $M=63\%$, 95% CI [59, 67]) and those for SI were relatively low (positive: $M=29\%$, 95% CI [25, 33]; negative: $M=25\%$, 95% CI [21, 29]). In line with Exp.1-2, these results suggest that FC and DI tend to be more robust than regular SIs. Crucially, however, no significant interactions between Condition and Polarity or between Inference type and Polarity was found. The model for the FC trials only yielded a main effect of Condition ($\chi^2(1) = 65.23$, $p < .001$) while the model for the **Target** conditions only yielded a main effect of Inference type ($\chi^2(1) = 44.15$, $p < .001$).⁶ No reliable contrast between positive and negative forms was found for any inference type (all β s < 3.71 , all p s > 0.1). These results establish that the low negative variants of FC and DI are just as robust as their positive forms.

3.7 Discussion

The contribution of this first set of experiments is twofold. First, the results reproduce previous findings from the literature regarding the strength of positive and high negative FC compared to regular SIs. Specifically, the results from Exp.1 replicate the main findings from Marty et al. 2021 (Exp.1-2; see also Chemla 2009) in showing that high negative sentences give rise to FC inferences but that these inferences are far less robust for these sentences than for their classical, positive counterparts. Importantly, in contrast to FC inferences, regular SIs were unaffected by our manipulation of sentence polarity in all three experiments that we carried out. This finding suggests that the large contrasts observed between positive and high negative FC cannot be attributed to a general difference in robustness due to the presence of negation.

Second, the present results extend and refine the previous findings from Marty et al. 2021 in at least two significant ways. To begin with, we found that the difference in strength observed between positive and high negative FC (Exp.1) extends to the comparisons involving the intermediate negative forms of these inferences (Exp.2), but not to those involving their low negative forms (Exp.3). This latter finding also suggests that what matters most is not the presence of negation *per se*. Next, we

⁶ On the initial optimization run, the optimizer stopped after a certain number of iterations without finding a solution. To increase the number of iterations, the algorithm was restarted from the (fitted) values that it had previously reached.

What makes an inference robust?

found that, as far as sentence polarity is concerned, distributive inferences pattern with FC inferences and distinctly from regular SIs. This is clearly established in our data by the fact that the high and intermediate negative variants of DI were also found to be far less robust than their positive forms (Exp.1-2) whereas their low negative variants were found to be just as robust (Exp.3).

Thus, of the three sets of comparisons that we carried out, only those between positive and low negative cases yielded results that are compatible with the Extended Alternative-based Hypothesis (see Table 2) in showing that low negative FC and DI are fairly robust and at least as robust as their positive counterparts. The comparisons between positive and high/intermediate negative cases, on the other hand, yielded results that refute the predictions of that hypothesis insofar as the expected robustness of negative FC and DI is concerned.

4 Experiments 4-6: Epistemic modals and ignorance

Adapting the covered box paradigm from Noveck (2001), Exp.4-6 investigated the epistemic equivalents of the positive and negative deontic sentences from Exp.1-3. Our primary goal was to test whether the main findings from Exp.1-3 extend to FC and DI cases involving epistemic modalities (i.e., *possible/certain*) in place of deontic ones (i.e., *permitted/required*).

In addition, Exp.4-6 also compare the epistemic instance of FC and DI to another well-known type of epistemic inferences associated with disjunctive sentences and generally referred to as IGNORANCE inferences (henceforth, II):

- (32) IGNORANCE INFERENCE (II)
- a. The box contains either a blue ball or a yellow ball.
 - b. \rightsquigarrow *The speaker does not know whether the box contains a blue ball and she does not know whether it contains a yellow ball*

Just like FC and DIs, the IIs arising from simple disjunctive sentences have been hypothesised to arise as SIs on the basis of non-lexical alternatives (a.o., Gazdar 1979, Sauerland 2004, Fox 2007, Meyer 2013, Fox 2016, Buccola & Haida 2019, Marty & Romoli 2021). While this idea has been couched in different frameworks over the years, all implicature-based approaches to IIs rely on the same two key assumptions.⁷ First, they assume that disjunction has its standard Boolean meaning and

⁷ There are two major implicature-based accounts of IGNORANCE inferences in the literature. The first account relies on a cooperativity principle akin to the Gricean Maxim of Quantity whereby what is actually said is compared to whatever else the speaker could have said that would have also been relevant (Gazdar 1979, Sauerland 2004, Fox 2007 among others). The second account relies on the idea that every utterance involves a silent modal operator, notated 'K', ranging over the speaker Doxastic's worlds (Meyer 2013, Fox 2016, Buccola & Haida 2019, Marty & Romoli 2021).

that IIs arise on the basis of the sentence’s literal meaning via scalar strengthening. Second, they assume that the IIs associated with a sentence like (32a) are derived on the basis of alternatives that are parts of the asserted sentence. Specifically, the comparison set is usually hypothesised to include (at least) the propositions corresponding to the disjuncts, leading to the conclusion that the speaker does not know which of the two disjuncts is true, thereby deriving the II in (32b).⁸

$$(33) \quad \left\{ \begin{array}{ll} \text{the box contains a blue ball or a yellow ball} & (a \vee b) \\ \text{the box contains a blue ball} & a \\ \text{the box contains a yellow ball} & b \\ \text{the box contains a blue ball and a yellow ball} & (a \wedge b) \end{array} \right\}$$

In sum, as in the case of FC and DI, the kind of alternatives over which IIs are derived do not involve any form of lexical substitution. The same observations hold of the high negative case of II in (34) and its low negative variant in (35) – the de-Morgan equivalent of (34) – both of which rely on the negative counterparts of *a* and *b* above, namely *the box does not contain a blue ball* ($\neg a$) and *the box does not contain a yellow ball* ($\neg b$).

(34) HIGH NEGATIVE II

- a. The box does not contain both a blue ball and a yellow ball.
- b. \rightsquigarrow *The speaker does not know whether the box contains a blue ball and she does not know whether it contains a yellow ball*

(35) LOW NEGATIVE II

- a. The box does not contain a blue ball or it does not contain a yellow ball.
- b. \rightsquigarrow *The speaker does not know whether the box contains a blue ball and she does not know whether it contains a yellow ball*

The differences between these two accounts are irrelevant to the present discussion since both of them make the same predictions for the cases and comparisons that we are interested in.

⁸ For completeness, we note that, on the grammatical account to IGNORANCE inference, the derivation of the II in (32b) involves two occurrences of EXH, one below and one above the matrix-K operator (Meyer 2013), as shown in (i). Thus, on this account, the alternatives needed to derive the inferences of interest correspond more specifically to the (exhaustified) disjuncts appearing in the scope of K.

(i) EXH[K[EXH[the box contains a blue ball or a yellow ball]]]

$$(ii) \quad \left\{ \begin{array}{ll} K[\text{EXH}[\text{the box contains a blue ball or a yellow ball}]] & K[(a \vee b) \wedge \neg(a \wedge b)] \\ K[\text{EXH}[\text{the box contains a blue ball}]] & K[a \wedge \neg b] \\ K[\text{EXH}[\text{the box contains a yellow ball}]] & K[b \wedge \neg a] \\ K[\text{EXH}[\text{the box contains a blue ball and a yellow ball}]] & K[a \wedge b] \end{array} \right\}$$

What makes an inference robust?

Once coupled with the Alternative-based hypothesis, the predictions of the implicature approach for FC and DI immediately extend to II too. That is, this approach predicts that (i) the IIs associated with simple disjunctive sentences should be very robust, unlike regular SIs, and (ii) their negative variants should behave in a similar way, regardless of whether negation appears at the matrix level or within each disjunct. These novel predictions, summarised in Table 6, were tested alongside those for FC and DI which we previously discussed, by carrying the same systematic comparisons as before.

Before going on, we note that we consider the inclusion of *IGNORANCE* inferences in this set of experiments as an exploratory first step. The reason is that the present design was not specifically geared at making salient the knowledge of the speaker, which is at the basis of these inferences. Nonetheless, we think that the results we report on below are suggestive and can serve as a basis for future investigations employing a design more suited to test this type of inference, as proposed for instance in Hochstein et al. 2016 and Dieuleveut, Chemla & Spector 2019, among others.

Environment	Inference type			
	FC	DI	II	SI
POSITIVE	✓	✓	✓	✓
NEGATIVE				
High	✓	✓	✓	✓
Intermediate	✓	✓	✓	
Low	✓	✓	✓	✓

Table 6 Predictions of the implicature approach, supplemented with the Alternative-based hypothesis. This table extends Table 2 so as to include the predictions for the novel II cases tested in Exp.4-6.

4.1 Participants

For each experiment, 70 novel participants were recruited online through Prolific using the same pre-screening criteria as in Exp.1-3. The recruitment was set up so that each participant could only take part in one of the three experiments. Participants were paid £1.75 and average completion time was around 12 minutes. The consent and data collection procedures were the same as in Exp.1-3 (see Section 3.1 for details).

4.2 Materials

All three experiments were designed in a similar way by adapting the sentences presented to participants to the specific purpose of each experiment, as in Exp.1-3. A schematic description of the sentence types tested in each experiment is given in Table 7. The different color adjectives used in the sentences, indicated by the [A] and [B] terms in Table 7, were randomly chosen among four possible options: *green, blue, yellow* or *grey*.

Each experiment tested four positive sentences, one for each inference type of interest (FC, DI, II and SI), and compared them to one of their negative variants. The different negative variants tested in Exp.4-6 were obtained from the positive sentences following the same logic as in Exp.1-3. The high negative cases, tested in Exp.4, were obtained by placing a negation at matrix level and by replacing the embedded modals and connectives with their scalemates (i.e., *possible*→*certain*, *certain*→*possible*, *or*→*and*). The intermediate negative cases for FC, DI and SI, tested in Exp.5, were obtained by placing a negation further down, at the level of the main verb of the embedded clause (i.e., *contain*→*not contain*), and by replacing, wherever relevant, the embedded connectives with their scalemate (i.e., *or*→*and*). The negative II sentences in Exp.5 were the same as in Exp.4. Finally, the low negative cases for FC, DI and II, tested in Exp.6, were constructed by placing a negation in each sub-clause of the complex embedded clause; the contents of these sentences were slightly adjusted to ensure that they sound as natural as possible to native speakers and, for the sake of parallelism, similar adjustments were applied to their positive counterparts in Exp.6. The negative SI sentences in Exp.6 were the same as in Exp.5.

Picture types and sentence-picture combinations were the same in all three experiments. Every picture displayed a quadruplet of boxes horizontally arranged (see examples in Figure 1). Each quadruplet was made of three open boxes, containing one or two balls, and a closed one, placed at the rightmost position, marked with the symbol "?" and referred to as "the mystery box". The contents of the open boxes on each picture were determined in reference to the sentence it was paired with, and their contents were experimentally manipulated so as to create true, false and target pictures for each sentence type, as described and illustrated in Table 8. Target pictures were designed so as to make the relevant sentence false if the inference of interest (i.e., FC, DI, II or SI) is present, but true if it is absent; by contrast, true and false pictures were designed so as to make the relevant sentence respectively true and false, independently of the inferences under scrutiny.⁹

⁹ In designing the pictures, further precautions were taken, on a case-by-case basis, to prevent other commonly observed implicatures from affecting participants' judgments. In particular, for the

What makes an inference robust?

Exp.4 – High negation

POSITIVE

- FC It is possible that the mystery box contains either a [A] ball or a [B] ball.
- DI It is certain that the mystery box contains either a [A] ball or a [B] ball.
- II The mystery box contains either a [A] ball or a [B] ball.
- SI It is possible that the mystery box contains a [A] ball.

NEGATIVE

- FC It is not certain that the mystery box contains both a [A] ball and a [B] ball.
- DI It is not possible that the mystery box contains both a [A] ball and a [B] ball.
- II The mystery box does not contain both a [A] ball and a [B] ball.
- SI It is not certain that the mystery box contains a [A] ball.

Exp.5 – Intermediate negation

POSITIVE

- FC It is possible that the mystery box contains either a [A] ball or a [B] ball.
- DI It is certain that the mystery box contains either a [A] ball or a [B] ball.
- II The mystery box contains either a [A] ball or a [B] ball.
- SI It is possible that the mystery box contains a [A] ball.

NEGATIVE

- FC It is possible that the mystery box does not contain both a [A] ball and a [B] ball.
- DI It is certain that the mystery box does not contain both a [A] ball and a [B] ball.
- II The mystery box does not contain both a [A] ball and a [B] ball.
- SI It is possible that the mystery box does not contain a [A] ball.

Exp.6 – Low negation

POSITIVE

- FC It is possible that either the mystery box contains a [A] ball or it contains a [B] ball.
- DI It is certain that either the mystery box contains a [A] ball or it contains a [B] ball.
- II Either the mystery box contains a [A] ball or it contains a [B] ball.
- SI It is possible that the mystery box contains a [A] ball.

NEGATIVE

- FC It is possible that either the mystery box does not contain a [A] ball or it does not contain a [B] ball.
- DI It is certain that either the mystery box does not contain a [A] ball or it does not contain a [B] ball.
- II Either the mystery box does not contain a [A] ball or it does not contain a [B] ball.
- SI It is possible that the mystery box does not contain a [A] ball.

Table 7 Schematic description of the sentences tested Exp.4-6 by sentence polarity and inference type, where [A] and [B] are placeholders for different colour adjectives; for a more concrete illustration, you may read [A] as *green* and [B] as *blue*.

positive and negative FC-sentences, all the pictures were designed so as to be compatible with the regular direct and indirect SIs that may arise from these sentences.

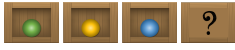
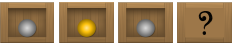

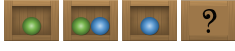


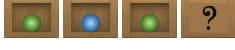


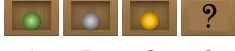


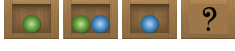
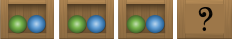
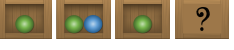
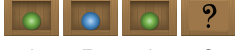
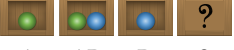







	True	False	Target
POSITIVE			
FC	 A C B ?	 D C D ?	 A C A ?
DI	 A AB B ?	 A C B ?	 A A A ?
II	 A B A ?	 D C D ?	 A A A ?
SI	 A D C ?	 D D D ?	 A A A ?
NEGATIVE			
FC	 A AB B ?	 AB AB AB ?	 A AB A ?
DI	 A B A ?	 A AB B ?	 A A A ?
II	 A B A ?	 AB AB AB ?	 A A A ?
SI	 A D A ?	 A A A ?	 D D D ?

Table 8 Schematic description and illustration of the picture types used in Exp.4-6. The color of the A-balls and B-balls always matched the color adjectives used in the sentence (e.g., *green* and *blue*) while the color of the C-balls and D-balls never did (e.g., *grey* and *yellow*). Picture types are illustrated here using the following color assignment: A=green, B=blue, C=yellow and D=grey.

The rest of the design was identical across all three experiments in all relevant aspects. In each experiment, each sentence type was paired with all three picture types, giving rise to the eponymous **True**, **False** and **Target** conditions. Each condition was iterated three times, resulting in 72 test trials. For each trial, the color adjectives used in the sentence was picked at random from our list of color terms, with replacement across trials. The color of the balls on the accompanying picture was determined according to the relevant sentence and the relevant condition: the colors of A-balls and B-balls always matched the color terms used in the sentence; the colors of the C-balls and D-balls were randomly chosen from our list by excluding the color(s) of the matching balls.

4.3 Procedure

The three experiments were run as online surveys using Gorilla Experiment Builder (Anwyl-Irvine et al. 2020). The procedure was the same in all three experiments. In the instructions, participants were introduced to two characters, Sam and Mia, and they were presented with a short cover story. The cover story was as follows (see Appendix B.1 for details): Sam and Mia are looking at quadruplets of boxes containing balls of various colors. For each quadruplet, they can only see what's inside the first three boxes. However, they know that the fourth box, known as "the mystery box", always has the same contents as one of the three open boxes and, therefore, they can make certain inferences about what's inside the mystery box. Participants were shown two example quadruplets where the contents of the mystery box were progressively revealed, allowing them to verify that its contents were indeed identical to those of one of the open boxes. Participants were told that they would see many quadruplets like these, each of which would be followed by an utterance from either Sam or Mia about what the mystery box contains, and that their task was to decide if this utterance is or is not a good description of what's inside the mystery box. They were instructed to click on 'Good' if they consider the character's utterance a good description of the picture and otherwise to click on 'Bad'.

Following the instructions, participants started the survey with a short training devised to consolidate their understanding of the cover story and instructions (see Appendix B.2 for details). After the training phase, the survey continued with a block of 72 test trials. Trials were presented in random order, with a 1000 ms interstimulus interval. Participants reported their judgments by clicking with the mouse one of two response buttons labelled 'Good' and 'Bad', respectively. The position of the response buttons (i.e., on the left or on the right) was counterbalanced across participants. Items remained on the screen until participants gave their response. At the end of the survey, participants were asked to fill out a short demographic questionnaire.

4.4 Data treatment

Data treatment was the same as in Exp.1-3. That is, responses from participants whose performance in the control trials did not reach the pre-established threshold of 70% accuracy were excluded prior to analyses. Responses from 4 participants in Exp.4 and from 5 participants in both Exp.5 and Exp.6 were removed through this procedure, representing about 5% and 7% of the data collected in these experiments, respectively.

4.5 Data analyses

Data was analysed using the data analysis pipelines from Exp.1-3 (see section 3.5) to address the following three questions. First, are the positive instances of FC and DI with epistemic modals more robust than direct SIs, as previously observed for FC and DI with deontic modals? Second, are these inferences as robust in all their negative forms as in their positive forms? Third, how do IIs fare compared to the three other inference types? Specifically, are these inferences affected by sentence polarity in a similar way as FC and DI?

To address these questions, the data from each experiment was analysed by carrying out two main analyses. In the first analysis, we assessed whether participants' responses in the **True** and in the **Target** conditions differ as a function of the inference type for both the positive and the negative sentences. The models included Condition (2 levels: True, Target; sum-coded), Inference type (4 levels: SI, FC, DI, II; dummy-coded) and their interaction as fixed effects, with Subject as a random effect and a by-Subject random slope for Condition.¹⁰ Inference type was coded with treatment contrasts using SI as a reference level, thus allowing us to use the regression output to compare SI to all three other inference types. In the second analysis, we assessed, for each of the four inference types, whether participants' responses in the **True** and in the **Target** conditions differ as a function of the sentence polarity. The models included Condition (2 levels: True, Target; sum-coded), Sentence polarity (2 levels: Positive, Negative; sum-coded) and their interaction as fixed effects, with the same random effect structures as the models previously described. As in Exp.1-3, wherever relevant, the main analyses were completed with post-hoc pairwise comparisons (p -values were adjusted for multiple testing, as before).

4.6 Results

The bar plot in Figure 4 shows the mean rates of rejection for each sentence type, in each experiment, by condition, inference type and sentence polarity. The interaction plot in Figure 3 shows in more details the mean rejection rates in the **Target** conditions for each sentence type as a function of sentence polarity.

4.6.1 Experiment 4: High negation

Exp. 4 compared the positive instances of FC, DI, II and SI to their high negative variants. For the positive cases, the mean rejection rates for FC, DI and II in the

¹⁰ This was the maximal random effect structure supported by the data. All models including Polarity as a random slope resulted in singular fits, indicating that these models were overfitted.

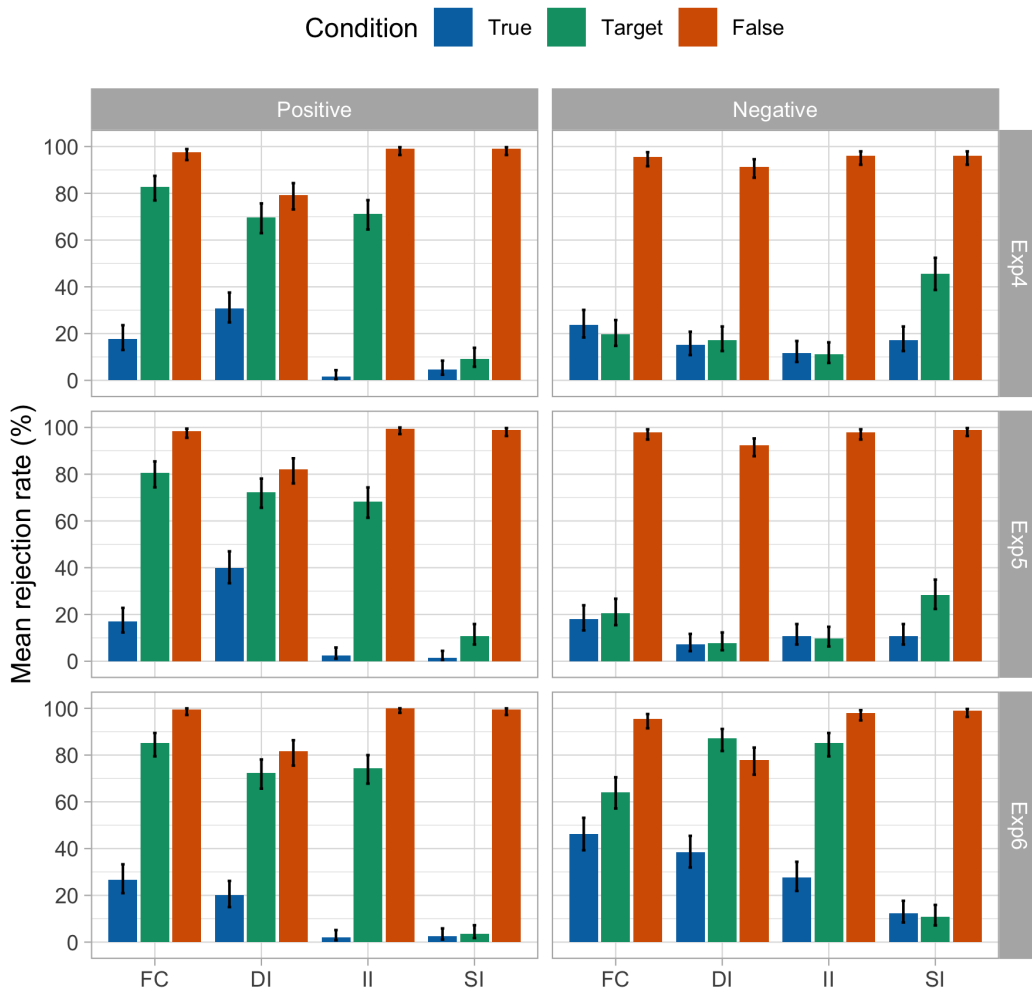


Figure 4 Mean rejection rate (in %) to each sentence type in Exp.4 (top), Exp.5 (middle) and Exp.6 (bottom) by sentence polarity and experimental condition. Error bars represent 95% confidence intervals.

Target conditions were all fairly high ($69\% < Ms < 83\%$) whereas that for SI was very low ($M = 9\%$).¹¹ For the negative cases, the mean rejection rates for FC, DI and II in the **Target** conditions were all fairly low ($20\% < Ms < 45\%$) whereas that for SI was somewhat intermediate ($M = 45\%$). The model for the positive sentences

¹¹ The mean rejection rate for the positive DI sentences in the **True** conditions was slightly higher than expected ($M = 30\%$, 95%CI [24, 38]). We believe, however, that this result could indicate that participants sometimes derived embedded SIs for the positive DI sentences, rendering these sentences false in their **True** conditions. We will come back to this point in the Discussion.

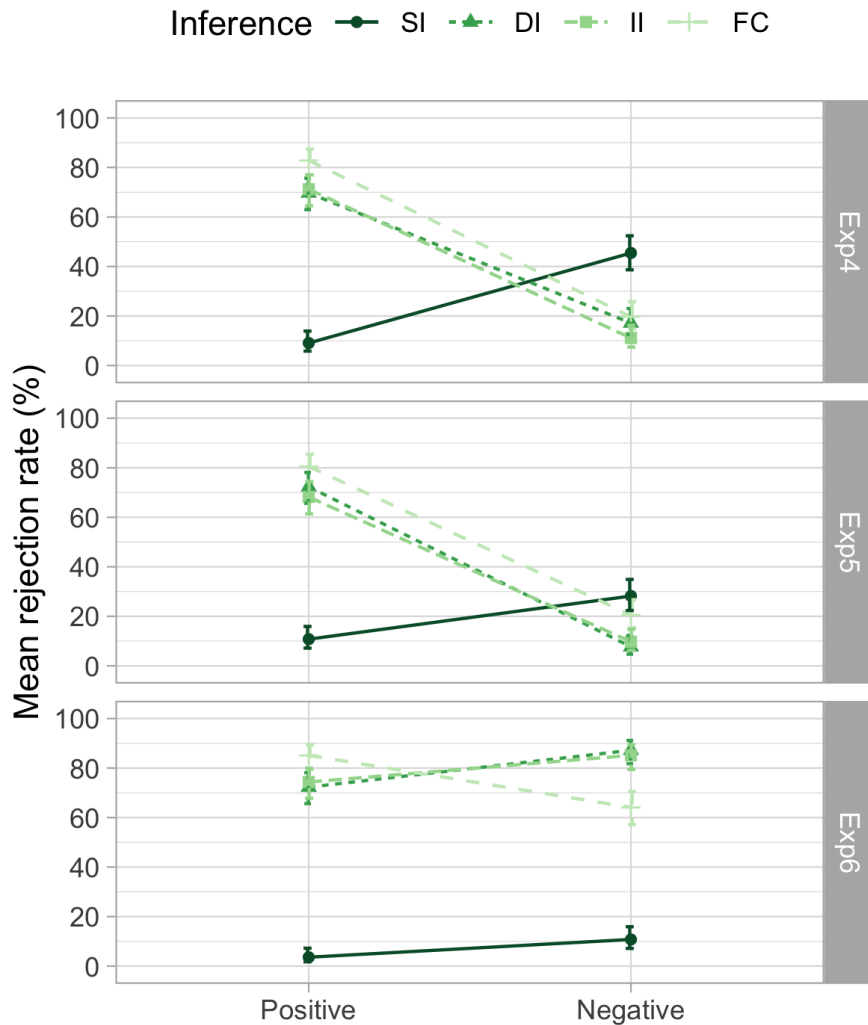


Figure 5 Mean rejection rates (in %) in the **Target** conditions of Exp.4 (top), Exp.5 (middle) and Exp.6 (bottom) by inference type and sentence polarity. Error bars represent 95% confidence intervals.

yielded a significant effect for Inference type ($\chi^2(3) = 155.95, p < .001$) and for the interaction between Condition and Inference type ($\chi^2(3) = 88.34, p < .001$) such that the difference between **True** and **Target** conditions was significantly smaller for SI than for DI ($\beta = 2.16, SE = 0.40, p < .001$), FC ($\beta = 3.38, SE = 0.44, p < .001$) and II ($\beta = 4.04, SE = 0.49, p < .001$). The model for the negative sentences yielded a significant effect for Condition ($\chi^2(3) = 38.63, p < .001$),

What makes an inference robust?

Inference type ($\chi^2(3) = 43.15, p < .001$) and their interaction ($\chi^2(3) = 28.24, p < .001$) such that, in contrast to what we found in the previous model, the difference between **True** and **Target** conditions was significantly larger for SI than for DI ($\beta = -0.65, SE = 0.18, p < .001$), II ($\beta = -0.74, SE = 0.20, p < .001$) and FC ($\beta = -0.88, SE = 0.17, p < .001$).

Turning to the second analysis, the models yielded a significant effect of Condition for each inference type (all $\chi^2_s > 6$, all $ps < .05$), a significant effect of Polarity for FC, DI and SI (all $\chi^2_s > 39$, all $ps < .001$; II: $\chi^2(1) = 1.49, ns$) and a significant interaction between both factors for FC, DI and II (all $\chi^2_s > 17$, all $ps < .001$; SI: $\chi^2(1) = 3.61, ns$) such that, for these three inference types, the difference between **True** and **Target** conditions was significantly larger for the positive sentences than for their negative counterparts. The effects of Condition were investigated further in a post-hoc analysis comparing the estimated marginal means for the **True** and **Target** conditions in the positive and in the negative cases: in the positive cases, there was a significant contrast between **True** and **Target** conditions for all inference types except SI; in the negative cases, the situation was reversed in that there was a significant contrast between both conditions only for SI.

Thus, the results from Exp.4 extend the results from Exp.1-3 in showing that, just like their deontic equivalents, FC and DI sentences involving epistemic modals readily gave rise to the eponymous inferences in their positive forms. The results also show that positive FC and positive DI were far more robust than positive SI and far more robust than their high negative variants. In fact, in contrast to what we found in our previous experiments, there is no evidence in the present data that participants accessed the inferences of interest for the latter, a striking difference we will get back to in the Discussion. Finally, the present results extend the current findings for FC and DI to II: all in all, FC, DI and II were found to pattern together, and distinctly from regular SIs, across all factors of interest.

4.6.2 Experiment 5: Intermediate negation

The comparisons between the positive and intermediate negative instances of FC, DI, II and SI yielded similar results as in Exp.4. The mean rejection rates were similar to those observed in Exp.4 across the board. The model for the positive sentences yielded a significant effect for Inference type ($\chi^2(3) = 165.98, p < .001$) and for the interaction between Condition and Inference type ($\chi^2(3) = 59.81, p < .001$) such that the difference between **True** and **Target** conditions was significantly smaller for SI than for II ($\beta = 2.43, SE = 0.46, p < .001$) and FC ($\beta = 1.87, SE = 0.44, p < .001$).¹² The model for negative sentences yielded a significant effect

¹² In contrast to what we found in Exp.4, the interaction involving positive SI and positive DI did not reach significance ($\beta = 0.62, SE = 0.43, p = .15$). A visual inspection of the results in Fig. 4

for Condition ($\chi^2(1) = 18.70, p < .05$), Inference type ($\chi^2(3) = 35.87, p < .001$) and their interaction ($\chi^2(3) = 13.49, p < .005$) such that the difference between **True** and **Target** conditions was significantly larger for SI than for II ($\beta = -0.72, SE = 0.24, p < .005$), FC ($\beta = -0.64, SE = 0.21, p < .005$) and DI ($\beta = -0.64, SE = 0.26, p < .05$). The results from the second analysis were also very similar to those reported in Exp.4. There was a significant effect of Condition for FC, DI and II ($\chi^2s > 7, ps < .005$), a significant effect of Polarity for each inference type (all $\chi^2s > 5$, all $ps < .05$) and a significant interaction between both factors for FC, DI and II ($\chi^2s > 9, ps < .005$; SI: $\chi^2(1) = 0.05, ns$). Finally, the results of the post-hoc analysis were the same as in Exp.4: there was a significant contrast between **True** and **Target** conditions for all inference types but SI in the positive cases, while there was no such contrast except for SI in the negative ones. In sum, the results from Exp.5 extend the main findings from Exp.4 to the comparisons between positive and intermediate negative cases.

4.6.3 Experiment 6: Low negation

The results from Exp.6 yielded different results compared to Exp.4-5 as far as the effects of sentences polarity are concerned. For the positive sentences, the results from Exp.6 were in line with those from Exp.4-5, suggesting that the slight rewording of these sentences in Exp.6 had little-to-no impact on how people interpreted them. The model for these sentences yielded a significant effect for Condition ($\chi^2(1) = 5.04, ps < .05$), Inference type ($\chi^2(3) > 112.45, ps < .001$) and their interaction ($\chi^2(3) = 50.40, p < .001$). As in Exp.4-5, the difference between **True** and **Target** conditions was significantly smaller for SI than for II ($\beta = 5.18, SE = 0.79, < .001$), FC ($\beta = 4.33, SE = 0.79, p < .001$) and DI ($\beta = 3.66, SE = 0.75, p < .001$). For the negative sentences, on the other hand, the results from Exp.6 stood in stark contrast with those from Exp.4-6: the mean rejection rates for negative FC, DI and II in the **Target** conditions were all relatively high ($64\% < Ms < 87\%$) whereas that for SI was very low ($M = 10\%$). The model for the negative sentences yielded a significant effect for Inference type ($\chi^2(3) > 237.21, ps < .001$) and the interaction between Condition and Inference type ($\chi^2(3) = 92.73, p < .001$) such that, as with the positive cases, the difference between **True** and **Target**

suggests that this difference is due to higher rejection rates for positive DI in the **True** conditions in Exp.5. This subtle variation between both experiments is immaterial for our purposes. We note, however, that one explanation for this variation could be that participants derived embedded SIs for positive DI sentences and did so to a different extent in Exp.4 and Exp.5 (see Discussion).

What makes an inference robust?

conditions was significantly smaller for SI than for II ($\beta = 1.84$, $SE = 0.22$, $p < .001$), DI ($\beta = 1.64$, $SE = 0.22$, $p < .001$) and FC ($\beta = 0.57$, $SE = 0.20$, $p < .01$).¹³

Turning to the second analysis, the models yielded a significant effect of Condition for FC, DI and II ($\chi^2_s > 13$, $ps < .001$; SI: $\chi^2(1) = 0.76$, *ns*), a significant effect of Polarity for DI, II and SI ($\chi^2_s > 20$, $ps < .001$; FC: $\chi^2(1) = 0.48$, *ns*) and significant interaction between both factors for FC, DI and II ($\chi^2_s > 4.5$, $ps < .05$; SI: $\chi^2(1) = 0.05$, *ns*) showing that, for these three inference types, the difference between **True** and **Target** conditions was significantly smaller for the negative than for the positive forms. The post-hoc analysis largely confirmed the general results from both analyses in showing that the contrasts between **True** and **Target** conditions were significant across the board for all inference types, except for SI.

Taken together, these results provide strong evidence for the existence of low negative FC, DI and II with epistemic modals and they demonstrate that, as far as sentence polarity is concerned, the low negative instances of these inferences are far more robust than regular SIs, in line with our previous conclusions from Exp.1-3 based on deontic modals. Finally, it bears pointing out that these inferences were detected using the same task and method as in Exp.4-5 and in the context of an experiment where, by contrast, there appears to be no sign of regular SIs, whether direct or indirect.¹⁴

4.7 Discussion

The results from Exp.4-6 extend the findings from Exp.1-3 in two significant ways. First, they extend the conclusions from Exp.1-3 to FC and DI sentences involving epistemic modals: just like their deontic equivalents, the positive and low negative forms of these sentences readily give rise to the inferences of interest whereas their high and intermediate negative forms do not. It is worth emphasizing that the various contrasts observed in Exp.1-3 between the positive and negative variants of these inferences were reproduced in Exp.4-6 using different stimuli and a different task. These results refute further the predictions of the Alternative-based Hypothesis (see Table 6).

Second, the results from Exp.4-6 show that the novel inference type that we tested in these experiments, namely II, patterned with FC and DI across all sets of comparisons. Specifically, we found that the mean rejection rates for the positive II

13 The original model for the negative sentences failed to converge, resulting in unreliable estimates and standard errors. To overcome this issue, this model was rerun with a simpler random effect structure obtained from the original one by removing the by-Subject random slope for Condition.

14 The fact that we didn't find any sign of regular SIs in this experiment is surprising as the SI sentences were the same as in Exp.5. While the source of these variations remains unclear to us at the moment, their presence invite us to interpret with caution the contrasts in inference strength observed in this experiment between SI and the other inference types in Exp.6.

sentences and their low negative variants were all relatively high and substantially higher than that for positive SI. By contrast, those for their high and intermediate negative variants were all fairly low and slightly lower than for negative SI. The caveat mentioned above about investigating this type of inferences in this design remains. In particular, in this design, we cannot exclude that the target sentences in simple disjunctive cases were rejected for reasons not directly related to ignorance. In particular, a description of a target situation in which all visible boxes contain a blue ball as ‘The mystery box contains a blue ball or a yellow ball’ could also be rejected on the basis of a maxim of manner-type reasoning leading to the conclusion that the speaker could have used the simpler alternative ‘The mystery box contains a blue ball.’ Having said this, for our purposes, what is important is that the same reasoning would apply to the negative case i.e. the speaker could have said ‘The mystery box does not contain a yellow ball’ rather than ‘The mystery box does not contain both a blue ball and a yellow ball.’ In other words, we acknowledge that at this stage there are open questions about the interpretation of the results relating to ignorance inferences, as well as how the reasoning above relates to a theory of such inferences. Nonetheless, the fact that the difference between positive and negative cases extends to II sentences across all comparisons of interest points to a similar theoretical challenge for theories of II as for theories of FC and DI.

Before turning to the general discussion, let us emphasise three aspects of the results of Exp.4-6 that we believe deserve further attention. The first one has to do with the fact that the implicature approach predicts the existence of negative FC, DI, and II inferences. As discussed, this prediction follows from the fact that the mechanisms giving rise to FC, DI and II in the positive cases should apply to their negative variants in a similar way, predicting the existence of comparable inferences. Given our linking hypothesis, this means that one could have expected the high and intermediate negative cases for FC, DI and II to be rejected more in the **Target** than in the **True** conditions, as we found for the positive and low negative cases. And this is indeed what Marty et al. 2021 found for high negative FC, which our results from Exp.1-3 replicate and extend to their intermediate negative variants. In other words, we take these results to provide solid evidence for the existence of negative FC.¹⁵ In Exp.4-5, on the other hand, we did not find a difference between **Target** and **True** conditions for any of the high and intermediate negative cases of FC, DI, and II. The fact that we did not detect the presence of these inferences in this study could indicate that some features of the task we devised made it more challenging for participants to engage in scalar reasoning, hence leaving

¹⁵ As mentioned above, given that we could not include **True** conditions for DI inferences in the design by Marty et al. 2021, we did not have the relevant comparison for testing the existence of high and intermediate negative DI inferences in Exp.1-3.

What makes an inference robust?

more room for literal interpretations.¹⁶ In support of this line of explanation, we note for instance that the rejection rates for positive SI in the **Target** conditions (about 10%) was lower than those found in Marty et al. 2021's studies (about 25%), evidencing that participants derived these inferences to a lesser extent in Exp.4-6. It is important to emphasise that this potential task-related effect, if present, should be general and, consequently, does not affect in any way our conclusions regarding the distribution of the sentence polarity effects.

The second point concerns an aspect of our results which we have already mentioned but not commented on yet (see fn.11 and fn.12). Specifically, we mentioned that the mean rejection rate for the positive DI sentences in the **True** conditions was slightly higher than expected in comparison to the other **True** positive conditions. As we did not design our experiments to probe for such differences, it is difficult to tell whether this observation should be given any theoretical importance in the context of our study. Nonetheless, we point out that this result could indicate that participants sometimes computed embedded SIs for these sentences, e.g., by applying the exhaustivity operator to each disjunct. That is, participants may have parsed $\Box(A \vee B)$ as $\Box(\text{EXH}(A) \vee \text{EXH}(B))$, generating the additional inference $\Box((A \wedge \neg B) \vee (B \wedge \neg A))$. Crucially, this inference was false in the **True** conditions for positive DI, as one of the three open boxes made $(A \wedge B)$ possible, while it is true in the corresponding **Target** conditions. While such embedded SIs may have also been derived for positive FC and DI sentences, it is worth emphasizing that these SIs were, by contrast, always true in the **True** and **Target** conditions associated with these sentences. Thus, the presence of embedded SIs could explain the circumscribed variations observed among the positive **True** conditions in Exp.4-6.¹⁷

The third and last point concerns the relative strength of the FC, DI and II inferences arising from the low negative cases compared to those arising from the positive ones. Taken at face value, the relatively high rejection rates observed in the **Target** conditions suggest that these inferences are quite strong. At the same time, participants also rejected to some extent the low negative variants of FC, DI and II in their **True** conditions; furthermore, the rejection rates for these sentences in their **True** conditions were substantially higher than those for their positive

16 It is possible, for instance, that the reasoning process required to infer the possible contents of the mystery box heavily drew on participants' working memory resources, affecting their pragmatic reasoning in a way similar to a dual memory task (De Neys & Schaeken 2007, Marty & Chemla 2013, Antoniou, Cummins & Katsos 2016, van Tiel, Pankratz & Sun 2019, van Tiel et al. 2019, Marty et al. 2020). For the time being, we simply note that the average decrease in mean rejection rates observed for positive SI in our study, compared to Marty et al. 2021's Exp.3, is in line with the decrease generally induced by higher memory load in dual task studies.

17 Similar discrepancies among the positive **True** conditions were found in Exp.4 and Exp.5, but not in Exp.6, where the structure of the DI sentences was slightly different.

versions in the corresponding conditions. These results suggest that participants' judgements for these sentences may have been affected by considerations beyond the evaluation of their truth and felicity. One such consideration could relate for instance to the naturalness of these sentences: compared to the negative sentences tested in Exp.4-5, the low negative sentences were lengthier (no phonological reduction) as well as structurally more complex (two negations instead of one), making them less natural than their high and intermediate variants. Thus, it is conceivable that some participants found these sentences quite cumbersome and used the "Bad" response option to report not only a judgment of falsity, but also a perceived defect of naturalness.

Setting the above explanation aside, the fact that, at a general level, participant's acceptance of the low negative FC, DI and II sentences were negatively impacted by independent factors has some consequences for the interpretation of our results. Essentially, it tells us that, for these sentences, the rejection rates observed in the **Target** conditions should not be taken at face value and should not be used in isolation to evaluate the strength of the relevant inferences. We note however that, regardless of the factor at play here, its effect should be general and affect participant's responses to these sentences in a similar way across experimental conditions. As a result, the outcome of the comparisons that we carried out, across sentence polarity and across inference type, can still be used to approximate the strength of the inferences at hand. With these considerations in mind, we conclude that our results suggest that low negative FC, DI and II are not as robust as their positive counterparts while they are clearly more robust than their high and intermediate variants and certainly more robust than high and low negative SI (see fn.14 for refinements).

5 General discussion

On the implicature approach, the derivation of FREE CHOICE, DISTRIBUTIVE and IGNORANCE inferences, in both their positive and negative instances, are assumed to be based on alternatives that do not involve lexical substitution. Our results are challenging for this uniform approach if it is supplemented with the Alternative-based hypothesis in (20), repeated below.

(20) **Alternative-based hypothesis (extended)**

Alternatives that do not involve lexical substitutions give rise to inferences that are more robust, faster to process and easier to acquire.

According to (20), inferences involving non-lexical alternatives should be more robust. The challenge for this hypothesis comes from the finding that the availability of FC, DI as well as II substantially differs across the various positive and negative

What makes an inference robust?

environments that we tested. Specifically, these inferences were either undetected or found to be relatively weak in the high and intermediate negatives cases, where negation scopes over conjunction, while they were systematically detected and found to be quite robust in the logically equivalent, low negative cases, where negation appears in the scope of disjunction. The challenge is strengthened by the fact that no such difference was found between the positive and negative instances of regular SIs; in fact, for these inferences, a contrast in the opposite direction was even observed in Exp.4-5. The findings from Exp.1-3 and Exp.4-6 are summarised in Table 9, alongside the predictions of the implicature approach supplemented with the Alternative-based hypothesis: these predictions are not empirically borne out for the high and intermediate negative cases. It is worth emphasising that the problem here is not *per se* that the relevant inferences were found to be relatively weak (Exp.1-3) or, in some cases, were not detected at all (Exp.4-6) in these cases. Rather, the problem comes from the differences observed between these cases and their positive and low negative counterparts, which were all computed at quite high rates across all our experiments.

	Predictions				Findings Exp.1-3			Findings Exp.4-6			
	FC	DI	II	SI	FC	DI	SI	FC	DI	II	SI
POSITIVE	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
NEGATIVE											
High	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗	✓
Intermediate	✓	✓			✓	✓		✗	✗		
Low	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓

Table 9 Summary of the predictions of the implicature approach supplemented with the Alternative-based hypothesis, together with the main findings from Exp.1-3 (deontic modals) and Exp.4-6 (epistemic modals). We use ✗ to indicate that the relevant inference was not detected in our data.

Before going on, let us briefly consider another alternative-related hypothesis recently suggested by Bar-Lev (2018) and Bar-Lev & Fox (2020) (see also Schulz 2019 and Romoli, Santorio & Wittenberg 2020). The hypothesis in question is given in (36).

- (36) **The EXH-NEG scalemate hypothesis**
 EXH and negation are lexical alternatives. As a result, substituting one for the other gives rise to additional alternatives.

The idea behind the formulation in (36) is that a sentence involving negation (or EXH) has more lexical alternatives than usually assumed. Specifically, such a sentence is hypothesised to have additional lexical alternatives derived by replacing negation with EXH and vice-versa. It is crucial to observe that, since these additional alternatives enter scalar reasoning, their presence may affect the calculation of the sets of ‘innocently excludable’ and ‘innocently includable’ alternatives and, in turn, block certain scalar inferences that would be expected otherwise to go through. The question for us is thus whether this hypothesis could account for the various contrasts we found in our experiments. This hypothesis can be shown to account for the difference between positive and high negative cases for FC and DI, while correctly predicting that these inference types should be robust in the low negative cases. However, it is not clear that it can help with the corresponding difference we found for II, and, most importantly, it incorrectly extends to regular SIs, predicting that negative SIs should be generally weaker than positive SIs. In sum, while this hypothesis can explain some of our results, it does not provide a full account of them. As it seems, what we need is an account that can distinguish (i) between the positive and negative cases of FC, DI and II, without extending to regular SIs, and (ii) between the high and intermediate negative cases on one side, and the low negative cases on the other. In the next section, we outline two promising directions to account for our results: one based on an additional hypothesis about relevance, and the other based on a non-implicature approach to the positive instances of FC, DI and II.

6 Two directions

6.1 A relevance-based approach

We start by sketching a route for the implicature approach. The proposal is meant to account for the difference between the cases involving disjunction, whether positive or negative, and the other negative cases involving conjunction. The proposal is based on two main ingredients – a notion of relevance and a stipulation about how disjunction makes its disjuncts relevant – which we will illustrate in turn. We also point to certain predictions and features of the approach which need to be tested and developed.

6.1.1 Relevance and disjunction

As commonly assumed, we take implicature computation to be constrained by relevance. Specifically, we assume that the implicature associated with a given alternative is derived only if that alternative is relevant. One common way to characterise the set of relevant alternatives to a given utterance is to define them

What makes an inference robust?

as those alternatives that appropriately ‘address’ the understood Question Under Discussion (QUD). Following previous literature, one way to model this notion of relevance is to assume that the understood QUD is associated with a partition of the context set, which corresponds to the set of complete answers to this question (Bennett 1979, Groenendijk & Stokhof 1984, Groenendijk, Janssen & Stokhof 1984, Heim 1994). On this view, the notion of relevant proposition can be defined as in (37), where Q is the partition associated with the QUD: a proposition is relevant if and only if it does not distinguish between two worlds within a cell of Q (a.o., Groenendijk & Stokhof 1984, Marty 2017, Romoli 2012, Marty & Romoli 2021).

(37) **Relevance:** A proposition p is relevant in a context c given a partition Q of c iff for any cell $q \in Q$ and any $w, w' \in q$, $p(w) = p(w')$.

The second ingredient comes from the observation that disjunctive statements are generally perceived as felicitous only if their disjuncts are relevant alternatives (for discussion, see Marty & Romoli 2021, Simons 2001, Fox 2007, Singh 2008, Fox & Katzir 2011). This felicity condition would be evidenced by examples like (38), from Simons 2001, where the two disjuncts are not discourse-related in any obvious way and the resulting disjunction is perceived as odd.

(38) ??There is dirt in the fuel line or it is raining in Tel-Aviv.

In light of the definition we gave in (37), one way to understand this felicity condition on disjunction is to assume that a disjunctive statement is likely to be understood as an answer to a QUD to which its disjuncts are also possible answers. To put it differently, a disjunctive statement is preferentially understood as an answer to a question which makes both its disjuncts relevant because its disjuncts are preferentially construed as relevant alternatives. This interpretive preference can be formulated as follows:

(39) **Relevance for Disjunctions**

A disjunction is likely understood as an answer to an active QUD which is also addressed by its single disjuncts. That is, whenever a disjunction is relevant, its disjuncts are likely relevant as well.

Note that it is sufficient for our purposes that the above principle be probabilistic in nature, rather than absolute. In fact, as Marty & Romoli (2021) and Simons (2001) discuss, there are situations where the general preference stated in (39) can be counteracted. It is so for instance if a disjunctive statement is understood as an answer to a ‘Yes-No’, polar question. For instance, in a conversation like (40), where a doctor is asking the patient about the symptoms of a particular disease, the partition associated with A’s question has only two cells: one in which B hasn’t

had either of persistent dry cough or high temperature, and one in which she has had at least one of them, possibly both. Given this partition, the whole disjunction is relevant, but neither of the independent disjuncts is (i.e., neither corresponds to a cell or a union of cells in the partition).¹⁸

- (40) A: Have you had a persistent dry cough or high temperature?
B: I have (had a persistent dry cough or high temperature).

With these two ingredients in place, we can now go back to our cases. The general result, on this approach, is that positive and low negative cases involving disjunction are expected to give rise to more robust inferences than their high and intermediate negative variants involving conjunction, consistent with our data.

6.1.2 Positive and low negative cases

Recall that the positive and low negative cases involve disjunction. Thus, in these cases, the principle in (39) makes it so that the alternatives involving the disjuncts – which are central to the derivation of FC, DI and II – are likely relevant and, as a result, the corresponding inferences – be it FC, DI, or II – should be readily derived. To illustrate, consider for instance the case of IGNORANCE inferences. According to (39), a simple disjunctive sentence like (41) is likely to be understood as an answer to a QUD which makes its disjuncts relevant, e.g., *What does the box contain?*. Therefore, the IGNORANCE inferences associated with an utterance of (41) are expected to be fairly robust.

- (41) The box contains a blue ball or a yellow ball.
a. The box contains a blue ball.
b. The box contains a yellow ball.

Similarly, the low negative variant of (41) in (42) is likely to be understood as an answer to a QUD which makes its disjuncts relevant (e.g., *What does the box do not contain?*).

- (42) The box does not contain a blue ball or it does not contain a yellow ball.
a. The box does not contain a blue ball.
b. The box does not contain a yellow ball.

Finally, since the principle in (39) applies to disjunctions in general, it naturally extends to the corresponding cases giving rise to FC and DI. For these embedded

¹⁸ Other cases where this preference has been found to be counteracted are examples like *Either Jane won, or I'm a monkey's uncle*, which Simons (2001) calls 'monkey's uncle' disjunctions.

What makes an inference robust?

cases, however, we note that one would have to say more about how the principle in (39) applies to embedding environments (e.g., the embedding modals in the case of FC and DI). For preliminary attempts along these lines, we refer the reader to Marty 2017 and Marty & Romoli 2021.

6.1.3 High and intermediate negative cases

The principle in (39) is silent about cases that do not involve disjunction. In particular, it leaves open the possibility that sentences like (43) be preferentially understood as answers to QUDs which do not make their independent conjuncts relevant (e.g., *Did the box contain a blue and a yellow ball?*). Hence, the relevance-based approach we sketched does not make any prediction regarding the strength of the inferences associated with such sentences. Similar observations hold for the FC and II cases involving embedded conjunction.

- (43) The box does not contain a blue ball and a yellow ball.
- a. The box does not contain a blue ball.
 - b. The box does not contain a yellow ball.

In sum, the principle in (39) only states that the disjuncts of a disjunctive sentence are preferentially understood as relevant alternatives, predicting the implicatures derived from these alternatives to be quite robust. It is thus compatible with the inferences arising from conjunctive sentences being comparatively much weaker.

6.1.4 A note on regular SIs

For the same reasons as above, the principle in (39) is also silent about regular SIs as long as disjunction is not involved. Hence, it is compatible with regular SIs being generally less robust than positive and low negative instances of FC, DI and II, as suggested by our results. It bears pointing out, however, that the present account predicts that the regular ‘exclusivity’ SI associated with disjunctive sentences should in principle be more robust than the regular ‘inclusivity’ SI associated with their negated, conjunctive variants. Concretely, it predicts that the exclusive SI in (44c) associated with (44a), derived from the conjunctive alternative in (44b), should be more robust than the inclusive SI in (45c) associated with (45a) and derived from the disjunctive alternative in (45b).

- (44) EXCLUSIVE SI
- a. The box contains either a blue ball or a yellow ball.
 - b. The box contains a blue ball and a yellow ball.
 - c. \rightsquigarrow *The box does not contain both a blue ball and a yellow ball*

- (45) INCLUSIVE SI
- a. The box does not contain both a blue ball and a yellow ball
 - b. The box contains either a blue ball or a yellow ball.
 - c. \rightsquigarrow *The box contains either a blue ball or a yellow ball*

This prediction obtains because, according to the principle in (39), if (44a) is relevant, then its disjuncts - namely *the box contains a blue ball* and *the box contains a yellow ball* - are likely relevant as well. Since the general assumption in the literature is that relevance is closed under conjunction (Katzir 2007, Fox & Katzir 2011), it is natural to think that the conjunctive alternative in (44a) should also be perceived as relevant. While we did not test the exclusive and inclusive SIs of disjunction, we suspect that the expected contrast does not hold, but we leave a more detailed investigation of this question for future work.¹⁹

6.1.5 Summary

The relevance-based approach relies on an assumption about the calculation of relevance for disjunction which, as we explained, captures previous observations regarding certain felicity conditions that disjunctive statements are subject to. This account predicts FC, DI and II to be robust whenever disjunction is involved, regardless of whether a modal or a negation is present as well, consistent with our results. All in all, we think that this direction is more promising than the one relying on the distinction between lexical and non-lexical alternatives, or the one based on the idea that negation and EXH are lexical alternatives to one another. In the following, we outline another direction to account for our data, a hybrid approach combining a non-implicature account of the positive and low negative cases, with an implicature approach for the other ones.

6.2 A hybrid approach

As discussed in Marty et al. 2021, a natural response to the challenge coming from their results, which we replicated and extended, is to give up a unified approach to positive and (high and intermediate) negative FC, DI, and II, combining a non-implicature account to the robust cases, and an implicature account to the weaker ones. This is a non-trivial move, but not a completely *ad hoc* one, because all non-implicature approaches to these inferences need to be combined with a theory of implicatures to account for regular SIs. The extra step required here would be a theory of implicatures which derives the negative versions of the inferences above

¹⁹ The results from Marty et al. 2020 suggest that, if anything, the contrast goes in the opposite direction as expected: inclusive SIs like (45c) tend to be more robust than exclusive SIs like (44c).

What makes an inference robust?

as an implicature. Finally, in order to make the right predictions about the relative strength of these inferences, the non-implicature approach should derive only the robust cases, being the positive and the low negative ones. In the following, we sketch such a hybrid approach. Before moving to that, let us also note that [Aloni \(2022\)](#) has put forward a different version of a hybrid approach, not based on implicatures, which can also account for [Marty et al.'s \(2021\)](#) and our results in full. We refer the reader to her paper for details.

6.2.1 The non-implicature part

A number of authors have put forward a way to derive FC as a semantic entailment ([Aloni 2003, 2007](#), [Goldstein 2019](#), [Rothschild & Yablo 2018](#), [Simons 2005](#), [Willer 2017](#), [Zimmerman 2000](#)). An appealing feature of this approach is that, since semantic entailments are generally robust, it immediately explains the robustness of FC. Furthermore, many of these theories are equipped with an optional mechanism to cancel FC, which is useful for capturing the observation that FC exhibits more nuanced robustness than what would be expected for an obligatory semantic entailment. Another notable feature of this approach is that many of its implementations are couched in a richer semantic system than classical semantics. The reason is that, to derive FC as an entailment, the possibility modal has to be able to access the meaning of each disjunct in a compositional fashion. Typically, Inquisitive Semantics (or more broadly, Alternative Semantics) is employed for this purpose. One feature of this framework that is particularly relevant for us is that negation has the effect of (re)creating classical meaning. As we shall see, this feature proves useful for understanding the contrasts unveiled by our results between, on one side, positive and low negative cases and, on the other, high and intermediate negative cases. Recently, [Cremers et al. \(2017\)](#) has also put forward a theory of IGNORANCE inferences that uses the same type of non-classical semantic system. Although this theory does not treat IIs as plain entailments, it shares certain features with the entailment approach to FC that are useful in accounting for our findings. As we will demonstrate, one can also develop an entailment approach to DISTRIBUTIVE inferences using the same semantic system (cf. [Simons 2005](#)).

Building on the previous works cited above, we present below a concrete implementation of the non-implicature approach to FC, DI and II, and discuss its predictions for the cases involving negation and conjunction that we tested in our experiments. After pointing out some shortcomings of this theory, we will discuss the possibility of augmenting it with scalar implicatures, following the suggestion in [Marty et al. 2021](#).

6.2.2 Free choice inferences

The entailment approach to FC is often couched in Inquisitive Semantics (among others Ciardelli, Linmin & Champollion 2018) or related frameworks. The reason is that these richer semantic systems make it possible to define a notion of alternatives that derive FC as an entailment. One should be keep in mind here that these alternatives are theoretically distinct from, but can co-exist with the alternatives that are used in scalar implicature computation. In order to avoid confusion, let us call the former *alternative possibilities*, and the latter *formal alternatives*. In our implementation, alternative possibilities are sets of possible worlds, while formal alternatives are linguistic expressions.

A key feature of these systems is that disjunction typically creates a set of alternative possibilities corresponding to their disjuncts. In order to formalise this idea, sentence meaning is consistently lifted to a set of propositions. That is, an atomic sentence is taken to denote a singleton set, as shown in (46).²⁰ For clarity, we will use boldface when we speak of a proposition qua a set of possible worlds.

$$(46) \quad \llbracket \mathbf{p} \rrbracket = \{\mathbf{p}\}$$

A disjunctive sentence simply denotes the union of the denotations of the disjuncts:

$$(47) \quad \llbracket \phi \text{ or } \psi \rrbracket = \llbracket \phi \rrbracket \cup \llbracket \psi \rrbracket$$

As a number of authors have pointed out (Aloni 2003, 2007, Simons 2005, Willer 2017, Goldstein 2019), a framework like this allows one to encode FC in the meaning of the possibility modal as in (48).²¹

$$(48) \quad \llbracket \text{possible}(\phi) \rrbracket = \{\bigcap_{\mathbf{p} \in \llbracket \phi \rrbracket} \diamond \mathbf{p}\}$$

By way of illustration, let us take a simple case with two atomic sentences, **a** and **b**, as disjuncts. The resulting set contains a single proposition that **a** and **b** are both possible:

$$(49) \quad \llbracket \text{possible}(\mathbf{a} \text{ or } \mathbf{b}) \rrbracket = \{\bigcap_{\mathbf{p} \in \llbracket \mathbf{a} \text{ or } \mathbf{b} \rrbracket} \diamond \mathbf{p}\} = \{\diamond \mathbf{a} \cap \diamond \mathbf{b}\}$$

In contrast to the implicature approach to FC, an entailment approach needs to say something about how DOUBLE PROHIBITION comes about. That is, it needs to

²⁰ This is different from the basic version of Inquisitive Semantics, known as InqB, where sentence denotations are always downward closed sets. This is because, as far as FC and other inferences are concerned, only the maximal elements (known as ‘alternatives’) in such sets matter.

²¹ For better readability, we write $\diamond \mathbf{p}$ for the proposition (qua a set of possible worlds) that the proposition **p** is possible, that is, $\diamond \mathbf{p} := \{w \mid \exists w' [wRw' \wedge w' \in \mathbf{p}]\}$, where R is the relevant accessibility relation.

What makes an inference robust?

account for the fact that the meaning of (50) is stronger than the negation of its FC reading.

(50) It is not possible that the mystery box contains a blue ball or a yellow ball.

In this system, negation is usually defined as shown in (51). As one can verify, the semantics of negation does not help derive DOUBLE PROHIBITION.

(51) $\llbracket \text{not } \phi \rrbracket = \{\overline{\cup \llbracket \phi \rrbracket}\}$

In order to account for DOUBLE PROHIBITION, one can follow Goldstein 2019 for instance and assume that the possibility modal has a homogeneity presupposition that either all alternatives of its prejacent are possible or none of them is. On this assumption, a negated possibility modal embedding disjunction is now defined and true when neither of the disjunct is possible (see Goldstein 2019 for details), hence deriving DOUBLE PROHIBITION.

6.2.3 Distributive inferences

As far as we are aware, there is no fully developed entailment approach to DISTRIBUTIVE inferences in the literature (cf. Simons 2005). However, the account above can be extended to derive DIs as plain entailments as well. Specifically, we can bake in DI in the meaning of the universal modal, as shown in (52).²²

(52) $\llbracket \text{certain}(\phi) \rrbracket = \{(\Box \cup \llbracket \phi \rrbracket) \cap (\bigcap_{\mathbf{p} \in \llbracket \phi \rrbracket} \Diamond \mathbf{p})\}$

Let us unpack this denotation. The first bit, $\Box \cup \llbracket \phi \rrbracket$, is essentially the classical meaning of the necessity modal which requires that, in every accessible possible world, ϕ be true one way or another. The second bit, $\bigcap_{\mathbf{p} \in \llbracket \phi \rrbracket} \Diamond \mathbf{p}$, is the distributive inference, which amounts to the conjunction of statements of the form ‘ \mathbf{p} is possible’, for each proper alternative possibility \mathbf{p} . Note that, as in the case of FC, this meaning will derive something too weak for the negation of a necessity statement. For instance, a sentence like (53) would be true if it was certain that the box contained, say, a blue ball, and impossible that it contained a yellow ball.

(53) It is not certain that the mystery box contains either a blue ball or a yellow ball.

²² For readability, we write $\Box \mathbf{p}$ for the proposition that \mathbf{p} is necessary, that is, $\Box \mathbf{p} := \{w \mid \forall w' [wRw' \rightarrow w' \in \mathbf{p}]\}$. We assume that grand intersection with an empty domain of propositions returns the set of all possible worlds, W .

One way of solving this issue is to elaborate on Goldstein (2019) and postulate a presupposition in the meaning of the necessity modal that strengthens the negative sentence appropriately. In this case, the required presupposition would be that either the necessity statement is true – i.e., the proposition $(\Box \cup \llbracket \phi \rrbracket) \cap (\bigcap_{p \in \llbracket \phi \rrbracket} \Diamond p)$ holds – or the proposition $\overline{\Box \cup \llbracket \phi \rrbracket}$ is true. The latter disjunct strengthens the meaning of a negated necessity sentence like (53), deriving its intuitive meaning.

6.2.4 Ignorance inferences

For IGNORANCE inferences, we will follow the proposal by Coppock & Brochhagen 2013 (see also Cremers et al. 2017). This proposal assumes the same semantic system as the one we are developing here and derives IGNORANCE inferences using an independent pragmatic principle. Specifically, Coppock & Brochhagen 2013 make use of the following maxim:

- (54) **Inquisitive sincerity**
 Don't utter an inquisitive sentence if you already know how to resolve the issue that it expresses.

The sincerity maxim in (54) is concerned with the issue (or inquisitive content) expressed by a given sentence, that is, with the issue that is resolved by the alternative possibilities that this sentence introduces. A sentence ϕ is called inquisitive just in case the issue it expresses is not trivially resolved by the information conveyed by ϕ itself, i.e., just in case $\llbracket \phi \rrbracket$ contains more than one maximal set of possible worlds. Typically, a disjunctive sentence like *The box contains a blue ball or a yellow ball* is inquisitive in this technical sense because (i) it generates multiple alternatives, call them **blue** and **yellow**, and (ii) the information it conveys is not sufficient to determine which of these propositions contains the actual world. Thus, if a speaker utters such a sentence, assuming that the sincerity maxim is in force between the interlocutors, we can infer that this speaker does not know which of **blue** or **yellow** is true. These inferences, together with the semantics in (47), derive the IGNORANCE inferences associated with disjunctive sentences.

One noteworthy consequence of this proposal is that it predicts that negated disjunction should not rise to IGNORANCE inferences. The reason is that, as per the definition in (51), negation gives rise to singleton sets of alternative possibilities and, as a result, a negative statement is always non-inquisitive, trivialising the sincerity maxim in (54). Finally, let us emphasise that, on this theory, IGNORANCE inferences are neither semantic entailments, nor scalar implicatures. Therefore, one can make auxiliary assumptions about the robustness of such inferences so as to reflect our experimental results.

What makes an inference robust?

6.2.5 High and intermediate negation

We have seen how the theory derives FC and DI as plain entailments and II as pragmatic inferences. We have also seen that, when negation is applied to such sentences, the results are essentially as in classical semantics. Specifically, for FC and DI, this is achieved by the presuppositions of modals; for II, this is due to the fact that negation is generally assumed to give rise to a single alternative possibility, no matter how many alternative possibilities its argument has. This property of negation is actually important for the present discussion. Let us show what the theory predicts for sentences containing negation over conjunction, starting with the high negative cases. Consider for instance the negative FC sentence below:

(55) It is not certain that the mystery box contains both a blue ball and a yellow ball.

In order to identify the predicted meaning for this sentence, we first have to define the meaning of conjunction. The standard way is to use point-wise intersection as follows:

$$(56) \quad \llbracket \phi \text{ and } \psi \rrbracket = \{ \mathbf{p} \cap \mathbf{q} \mid \mathbf{p} \in \llbracket \phi \rrbracket \wedge \mathbf{q} \in \llbracket \psi \rrbracket \}$$

Note that, according to this definition, conjunction does not create new alternative possibilities.²³

Consider now a sentence of the form *certain*(ϕ and ψ). As discussed, this sentence will have a presupposition that either it is necessary that both ϕ and ψ (plus any distributive inference it may have, see fn. 23), or the conjunction of ϕ and ψ is not necessary. Keeping this in mind, consider now the negation of this sentence, the denotation of which is given in (57).

$$(57) \quad \llbracket \text{not}(\text{certain}(\phi \text{ and } \psi)) \rrbracket = \{ (\Box \cup \llbracket \phi \text{ and } \psi \rrbracket) \cap (\bigcap_{\mathbf{p} \in \llbracket \phi \text{ and } \psi \rrbracket} \Diamond \mathbf{p}) \}$$

Due to presupposition projection, this sentence should inherit the same presupposition as its positive counterpart. Factoring this presupposition in, the sole proposition in the above set can be simplified to: $(\Box \cup \llbracket \phi \text{ and } \psi \rrbracket)$. Importantly, note that this proposition does *not* entail negative FREE CHOICE since it comprises all worlds in which it is not the case that both ϕ and ψ are certain and, at the same time, includes worlds where either one of them is certain, if the other one is not certain. Hence, the present system predicts that positive sentences of the form *possible*(ϕ

²³ At the same time, a conjunction can inherit alternative possibilities in case the conjuncts already contain alternatives. Given that a conjunction of disjunctive statements (e.g. *It is possible that the mystery box contains a blue or yellow ball and a green or red ball*) generally give rise to FC, DI and II, this is a good feature of the account.

or ψ) should entail robust FC inferences, while negative sentences of the form *not(certain(ϕ and ψ))* should not. In light of our findings, and in particular the difference between positive and negative cases, this prediction appears correct. For similar reasons, the system does not derive a negative DISTRIBUTIVE inference as an entailment. Consider the sentence below:

- (58) It is not possible that the mystery box contains both a blue ball and a yellow ball.

This sentence is predicted to be true even if one of the most embedded conjuncts is certain and the other is impossible. To see this more clearly, consider a sentence of the form *not(possible(ϕ and ψ))*, which will denote the singleton set $\{\overline{\bigcap_{\mathbf{p} \in \llbracket \phi \text{ and } \psi \rrbracket} \Diamond \mathbf{p}}\}$. Recall that we follow Goldstein (2019) and assume that a possibility modal has a homogeneity presupposition. When this presupposition is taken into consideration, this denotation gets strengthened to $\{\bigcap_{\mathbf{p} \in \llbracket \phi \text{ and } \psi \rrbracket} \overline{\Diamond \mathbf{p}}\}$, i.e., none of the alternative possibilities are possible. Since each alternative possibility \mathbf{p} is a set of possible worlds where ϕ and ψ are both true, this amounts to saying that it's impossible that ϕ and ψ be simultaneously true, which does not exclude the situation in which one of ϕ and ψ is certain and the other impossible. Hence, negative DISTRIBUTIVE inferences are not derivable as plain entailments.

As we have already remarked, negative statements are generally non-inquisitive, and non-inquisitive sentences trivially satisfy the sincerity maxim. Hence, when negation takes scope over conjunction, as in (59), no IGNORANCE inference is predicted.

- (59) The mystery box does not contain both a blue ball and a yellow ball.

Let us now turn to cases where negation takes scope below the modal but above the connective, starting with the intermediate negative case for FC:

- (60) It is possible that the box does not contain both a blue ball and a yellow ball.

Recall that negation always returns a singleton alternative possibility and that, with a single alternative possibility, the possibility modal remains essentially classical since, in the absence of multiple alternative possibilities, the additional meaning component that normally derives FC has nothing to add. Consequently, the predicted meaning of (60) is simply classical, meaning that there is no FC inference in the semantics of this sentence.

Let us now turn to DI for a sentence with negation in an intermediate position, as in (61).

What makes an inference robust?

(61) It is certain that the box does not contain both a blue ball and a yellow ball.

Due to the negation that sits immediately below it, the necessity modal only has one alternative possibility to operate on, which means that the meaning stays essentially classical. Therefore, the sentence is predicted to have no distributivity inference as entailment.

6.2.6 Low negation

What about cases where negation appears below the connective? For these sentences, since disjunction is involved and takes scope over negation, the inferences should re-emerge. Consider first the low negative FC sentence below:

(62) It is possible that the mystery box either does not contain a blue ball or it does not contain a yellow ball.

The prediction here is the same as for the classical, positive instances of FC. That is, the sentence is predicted to entail that the alternative possibilities denoted by the disjunction are both possible. Concretely, (62) is predicted to entail that it is possible that the box does not contain a blue ball and that it is possible that it does not contain a yellow ball.

The same applies to the low negative cases for DI. The necessity modal combines with the alternatives of the disjunctive prejacent and requires each of them to be not certain. The resulting meaning for a sentence like (63) is thus as follows: it is certain that the box does not contain a blue ball or that it does not contain a yellow ball, and it is not certain that it does not contain a blue ball and it is not certain that it does not contain a yellow ball (i.e., it is possible that it contains one and it is possible that it contains the other).

(63) It is certain that the mystery box either does not contain a blue ball or it does not contain a yellow ball.

Finally, low negative II sentences like (64) generate similar alternative possibilities to their positive counterparts. Therefore, the sincerity maxim applies to these cases in the same way as before, giving rise to the expected IIs: the speaker does not know whether the box contains a blue ball and she does not know whether it contains a yellow ball.

(64) The mystery box does not contain a blue ball or it does not contain a yellow ball.

6.2.7 Summary

The non-implicature approach to FC, DI and II that we developed in this section derives all three types of inferences for the cases involving disjunction, whether negation is absent (positive) or takes scope below disjunction (low negative). On the other hand, for cases involving negated conjunction (high and intermediate negative), none of these inference types are derived as plain entailments. These predictions are consistent with our experimental results, but, in light of the results of Exp.1 and Exp.2, something more needs to be said to account for the availability of the negative inferences in the high and intermediate cases. As mentioned, in order to fully account for our results, the non-implicature approach can be combined with a theory of scalar implicatures to achieve this purpose. The resulting hybrid approach covers all of the cases that we investigated in our experiments.

It worth pointing out, however, that the hybrid approach raises questions about how the two types of alternatives it includes – formal alternatives and alternative possibilities – would interact with one another. To illustrate, consider for instance the modalised disjunctive sentence in (65), where a scalar term (here, *some*) is embedded in one of the disjuncts.

- (65) John might have met some of the students this morning or prepared next week's teaching.

This sentence appears to have a reading conveying that *John might have met some but not all of the students* and that *he might have prepared next week's teaching*, that is, a reading including both FC and the regular 'not-all' SI associated with *some*. Is this 'not-all' SI computed locally here? Do the formal alternatives associated with *some* interact with the alternative possibilities introduced by the disjunction? And, finally, what about the formal conjunctive alternative to disjunction, from which one can derive the exclusivity implicature that *it's not possible that John both met some of the students and prepared next week's teaching*? We leave these questions for future work exploring the underpinnings of the hybrid approach and its consequences.

7 Conclusion

Sentences involving embedded disjunction have long been observed to give rise to robust FREE CHOICE and DISTRIBUTIVE inferences. These two inferences exhibit strong similarities with regular SIs and, for this reason, a prominent take in the literature is to treat them as such. This uniform approach has been recently challenged by experimental findings showing a variety of differences between these inferences and regular SIs. The general response in the literature has been to explain these

What makes an inference robust?

differences in reference to the type of alternatives involved in the derivation of these inferences. This hypothesis, repeated below, relies on the fact that, on the implicature approach, the derivation of regular SIs involves formal alternatives which are derived by lexical substitution, while the derivation of FC and DI involve formal alternatives which are already part of the asserted sentence.

(20) **Alternative-based hypothesis (extended)**

Alternatives that do not involve lexical substitutions give rise to inferences that are more robust, faster to process and easier to acquire.

In this paper, we reported on a series of experiments testing this hypothesis by looking at other, related constructions for which the implicature approach predicts similar FREE CHOICE and DISTRIBUTIVE inferences on the basis of the same type of non-lexical alternatives. These constructions were all variants of the classical, positive cases which involved negation. We systematically compared the positive and negative instances of these inference types as arising and compared them to positive and negative instances of regular SIs. We further manipulated the position of negation to create additional logically equivalent cases, for free choice as well as for the distributive cases. In the first set of experiments, we investigated these inferences as arising from sentences involving deontic modals, building on and extending the study in [Marty et al. 2021](#). In the second set of experiments, we replicated the same set of comparisons using epistemic modals, adopting the mystery box design from [Noveck 2001](#). In addition, we extended the investigation to the IGNORANCE inferences (II) of simple disjunctive cases.

Our results across all experiments revealed large differences in robustness between the different cases. Specifically, we found that FC, DI and II were quite robust in cases involving disjunction, whether positive or negative. On the other hand, we found no trace of these inferences in cases involving negated conjunction, regardless of the position of negation. Furthermore, we found no corresponding difference between regular SIs and their negative counterparts. As we discussed, these findings are challenging for the idea that the implicature approach should be supplemented with the hypothesis in (20).

We outlined two directions to account for our results. The first one retains an implicature-based approach to FC, DI and II, but supplements it with an additional assumption about the calculation of relevance for disjunctive sentences. The second relies instead on a non-implicature approach, combined with an implicature one. Both approaches predict robust inferences for the positive and low negative cases – those cases involving disjunction – but not for the other negative cases involving conjunction, consistent with our results.

We end this discussion with two final remarks. First, while our results challenge the extended version of the Alternative-based hypothesis, they do not exclude its restricted version, repeated below, which restricts its scope of application to processing and acquisition. In other words, it could still be that the reason why these inferences are faster to process and easier to acquire than regular SIs is that their alternatives do not involve lexical substitutions. However, we take our results to show that their robustness does not stem from this property.

(19) Alternative-based hypothesis (restricted)

Alternatives that do not involve lexical substitutions give rise to inferences that are faster to process and easier to acquire.

One may wonder why (19) would hold while (20) does not. We think this suggests that the relationship between the salience of an alternative and the robustness of the corresponding inference is only indirect. We take this option to point to a psycholinguistic model of alternative-based inferences where the properties of an alternative, like its salience in virtue of being part of the asserted sentence, can reduce the processing cost and acquisition complexity of the resulting inference, but where the actual computation of this inference depends on further considerations such as the evaluation of whether or not this inference is intended in the first place. This model is very much in line with the relevance-based approach that we outlined, where the decision to compute scalar implicatures is mediated by the evaluation of the alternatives' relevance, even if these alternatives are active enough. It is less clear to us, at this stage, how the hybrid approach would fit with such a model. Be as it may, the next step is to investigate the processing and acquisition of the negative versions of FC, DI and II, and test whether they are processed similarly fast and acquired similarly early as their positive counterparts, as the hypothesis in (19) predicts.

Finally, we briefly note that the present study has close connections to some recent work on counterfactuals embedding disjunctive and negated conjunctive sentences (see in particular Romoli, Santorio & Wittenberg 2020, Ciardelli, Linmin & Champollion 2018, Schulz 2019). In fact, one of the hypotheses we discussed, the EXH-NEG scalemate hypothesis in (36), has been developed in response to the finding that disjunctive and negated conjunctive sentences embedded in counterfactuals are interpreted differently by speakers (despite being logically equivalent). Our results, through the differences we found between disjunctions and negated conjunctions, echo the results reported in that literature. An important next step is now comparing in detail the results between the simple and modalised cases we investigated and those involving counterfactuals, as well as testing further the predictions of each approach extended to both of these linguistic environments.

What makes an inference robust?

A Training trials and Instructions in Exp.1-3

A.1 Instructions for Exp.1-3

GENERAL INSTRUCTIONS

In this study, we will ask for your judgments about certain kinds of sentences in English. These sentences will involve two children, Sam and Mia. Here they are:



Sam



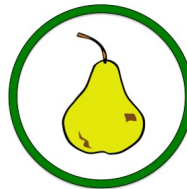
Mia

.....

Sam and Mia are going to the supermarket. Their parents have some rules about what Sam and Mia are allowed to buy, what they have to buy, and what they are not allowed to buy. Here is how we'll represent these rules. Take a look at the examples below:



The **red circle** around the burger means that *it is not permitted* to buy the burger.



The **green circle** around the pear means that *it is permitted but not required* to buy the pear.



The **black square** around the banana means that *it is required* to buy the banana.

.....

You will see many pictures depicting such rules. Each picture will be followed by a sentence that relates to it. Your task is to decide if that sentence is or not a good description of the picture you see. You will click on 'Good' if you consider the sentence a good description of the picture; otherwise click on 'Not good'.

TRAINING PHASE

You will start with a short training to get you familiar with the three basic rules that we have just introduced. During this training, you will receive **feedback** on your responses. If you answered correctly, you will see a **green smiley face** 😊; otherwise, you will see a **red frowning face** 😞 and be asked to try again. Use this feedback wisely to improve your answers.

We are interested in your spontaneous responses, so please don't think too long before answering.

TEST PHASE

From now on, the pictures you will see will depict rules about **two items**. As before, each picture will be followed by a sentence that relates to it and you will decide whether that sentence is or is not a good description of the picture you see. **You will no longer receive feedback on your responses.**

Recall that we are interested in your spontaneous responses, so don't think too long before answering.

A.2 Training trials in Exp.1-3

All participants in Exp.1-3 started with a first block of training trials. The purpose of this training phase was to ensure that participants got the instructions right and, specifically, that they understood correctly how to interpret in isolation the three symbols enclosing the food items. Pictures in these trials involved only one food item, enclosed within one of the three symbols. The training sentences were simple sentences involving similar lexical materials as the test sentences: there were two positive sentences of the form $\diamond(A)$ and $\square(A)$, and two high negative sentences of the form $\neg\diamond(A)$ and $\neg\square(A)$, as shown in (66). The [name] term was the name of one of the two characters and the [food] term was the name of the food item displayed on the picture.

- (66)
- a. It is permitted that [name] buys the [food].
 - b. It is not permitted that [name] buys the [food].
 - c. It is required that [name] buys the [food].
 - d. It is not required that [name] buys the [food].

Training sentences were presented with pictures that made them clearly true or clearly false. In a nutshell, sentences like (66a) and (66b) were paired with A and D pictures, and sentences like (66c) and (66d) were paired with A and R pictures. Each trial type was repeated twice. The result was a set of 16 training trials designed so that the proportion of expected 'Good' and 'Not good' responses was well-balanced within and between sentence type. Throughout the training phase, participants received feedback on the accuracy of their responses (see Appendix A.1 above).

B Training trials and Instructions in Exp.4-6

B.1 Instructions for Exp.4-6

GENERAL INSTRUCTIONS

What makes an inference robust?

In this study, we will ask for your intuitions about certain kinds of sentences in English. These sentences will be uttered by two characters, Sam and Mia. Here they are:



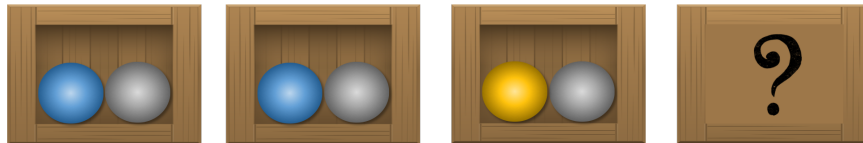
Sam



Mia

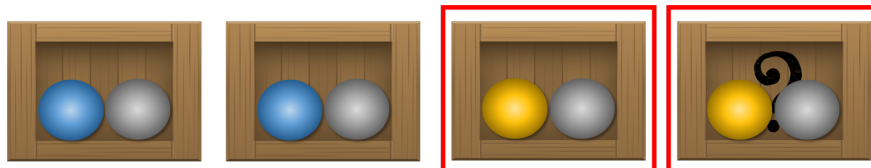
.....
What's inside the mystery box?

Sam and Mia are looking at four boxes containing balls of various colors. Each time, they can only see what's inside the first three boxes, as exemplified below. However, they know that **the fourth box always has the same contents as one of the three open boxes**, and therefore they can make certain inferences about what's inside this mystery box.



In this case, we can infer for instance that the mystery box must contain a grey ball, and that it can also contain either a blue ball or a yellow ball. Similarly, we can infer that it cannot contain a green ball, and that it cannot contain both a blue ball and a yellow ball.

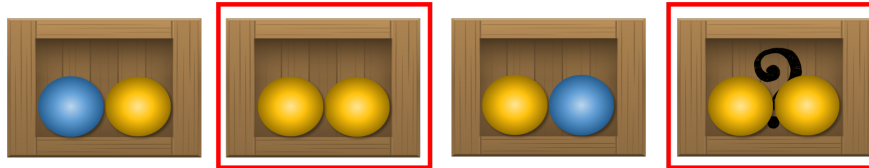
.....
In this case, the contents of the mystery box were identical to those of the third box.



.....
In this second example, we can infer for instance that the mystery box must contain at least one yellow ball. Similarly, we can infer that it need not contain both a blue ball and a yellow one, and that it need not contain a blue ball, etc.



.....
In this case, the contents of the mystery box were identical to those of the second box.



.....
You will see many quadruplets like these ones, each of which will be followed by an utterance from either Sam or Mia about what the mystery box contains. Your task is to decide if this utterance is or is not a good description of what's inside the mystery box. You will click on 'Good' if you consider the utterance a good description of the mystery box's contents; otherwise click on 'Bad'.

TRAINING PHASE

You will start with a short training to get you familiar with the study. During this training, you will receive **feedback** on your responses. If you answered correctly, you will see a **green smiley face** 😊; otherwise, you will see a **red frowning face** 😞 and be asked to try again. Use this feedback wisely to improve your answers.

We are interested in your spontaneous responses, so please don't think too long before answering.

TEST PHASE

As in the training, Sam and Mia will produce utterances about what the mystery box contains and you will decide whether these utterances are appropriate descriptions of the pictures you see. From now on, **you will no longer receive feedback on your responses.**

Recall that we are interested in your spontaneous responses, so don't think too long before answering.

B.2 Training trials in Exp.4-6

All participants in Exp.4-6 started with a first block of training trials. The purpose of this training phase was to ensure that participants got the instructions right and, specifically, that they understood how the possible contents of the mystery box could be inferred from the contents of the open boxes. The training sentences were distinct from the target sentences tested in each experiment, but they involved

What makes an inference robust?

similar lexical materials: there were two positive sentences of the form $\Box(A \wedge B)$ and $\Box(A)$, and two high negative sentences of the form $\neg\Diamond(A \vee B)$ and $\neg\Diamond(A)$, involving the modals *certain* and *possible*, as shown in (67).

- (67) a. It is certain that the mystery box contains both a [A] ball and a [B] ball.
b. It is certain that the mystery box contains a [A] ball.
c. It is not possible that the mystery box contains either a [A] ball or a [B] ball.
d. It is not possible that the mystery box contains a [A] ball.

Training sentences were presented with pictures that made them either clearly true or clearly false. Thus for instance, in the ‘Good’ training trials, sentences like (67b) were paired with pictures on which all the boxes contained an A-ball whereas, in the ‘Bad’ training trials, they were paired with pictures on which all but one box contained an A-ball. Each trial type was repeated 3 times. The result was a set of 24 training trials designed so that the proportion of expected ‘Good’ and ‘Bad’ responses was well-balanced within and between sentence type. Throughout the training phase, participants received feedback on the accuracy of their responses (see Appendix B.1 above).

References

- Aloni, Maria. 2003. Free choice in modal contexts. In *Proceedings of sinn und bedeutung* 7, 25–37.
- Aloni, Maria. 2007. Free choice, modals, and imperatives. *Natural Language Semantics* 15(1). 65–94. <https://doi.org/10.1007/s11050-007-9010-2>.
- Aloni, Maria. 2022. Logic and conversation: the case of free choice. *Semantics & Pragmatics* 15. <https://doi.org/doi.org/10.3765/sp.15.5>.
- Antoniou, Kyriakos, Chris Cummins & Napoleon Katsos. 2016. Why only some adults reject under-informative utterances. *Journal of Pragmatics* 99. 78–95.
- Anwyl-Irvine, A. L., J. Massonie, A. Flitton, N. Z. Kirkham & J. K. Evershed. 2020. Gorilla in our midst: an online behavioral experiment builder. *Behavior Research Methods* 52. 388–407.
- Bar-Lev, Moshe. 2018. *Free choice, homogeneity and innocent inclusion*. The Hebrew University of Jerusalem dissertation.
- Bar-Lev, Moshe & Danny Fox. 2020. Free choice, simplification, and Innocent Inclusion. *Natural Language Semantics* 28. 175–223. <https://doi.org/10.1007/s11050-020-09162-y>.
- Barner, David, N. Brooks & Alan Bale. 2011. Accessing the unsaid: the role of scalar alternatives in children’s pragmatic inferences. *Cognition* 188. 87–96.

- Bates, Douglas, Martin Mächler, Ben Bolker & Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1). 1–48. <https://doi.org/10.18637/jss.v067.i01>.
- Bennett, Paul A. 1979. On Universal 23. *Linguistic Inquiry* 10(3). 510–511.
- Bott, Lewis & Ira Noveck. 2004. Some utterances are underinformative. *Journal of Memory and Language* 51(3). 437–457. <https://doi.org/10.1016/j.jml.2004.05.006>.
- Breheny, Richard, Nathan Klinedinst, Jacopo Romoli & Yasutada Sudo. 2018. The symmetry problem: current theories and prospects. *Natural Language Semantics* 26(2). 85–110. <https://doi.org/10.1007/s11050-017-9141-z>.
- Buccola, Brian & Andreas Haida. 2019. Obligatory irrelevance and the computation of ignorance inferences. *Journal of Semantics* 36(4). 583–616.
- Chemla, Emmanuel. 2009. Universal implicatures and free choice effects: experimental data. *Semantics and Pragmatics* 2(2). 1–33. <https://doi.org/10.3765/sp.2.2>.
- Chemla, Emmanuel. 2010. Similarity: towards a unified account of scalar implicatures, free choice permission and presupposition projection. Ms., ENS.
- Chemla, Emmanuel & Lewis Bott. 2014. Processing inferences at the semantics/pragmatics frontier: disjunctions and free choice. *Cognition* 130(3). 380–396.
- Chierchia, Gennaro. 2013. *Logic in grammar: polarity, free choice, and intervention*. Oxford University Press.
- Chierchia, Gennaro, Danny Fox & Benjamin Spector. 2012. The grammatical view of scalar implicatures and the relationship between semantics and pragmatics. In Claudia Maienborn, Klaus von Stechow & Paul Portner (eds.), *Semantics: an international handbook of natural language meaning*, vol. 3. Berlin: Mouton de Gruyter.
- Ciardelli, Ivano, Zhang Linmin & Lucas Champollion. 2018. Two switches in the theory of counterfactuals: a study of truth conditionality and minimal change. *Linguistic and Philosophy* 41(6). 577–621. <https://doi.org/10.1007/s10988-018-9232-4>.
- Coppock, Liz & Thomas Brochhagen. 2013. Raising and resolving issues with scalar modifiers. *Semantics and Pragmatics* 3(6). 1–57.
- Cremers, Alexandre, Elizabeth Coppock, Jakub Dotlacil & Floris Roelofsen. 2017. Modified numerals: two routes to ignorance.
- Crnič, Luka, Emmanuel Chemla & Danny Fox. 2015. Scalar implicatures of embedded disjunction. *Natural Language Semantics* 23(4). 271–305. <https://doi.org/10.1007/s11050-015-9116-x>.
- De Neys, Wim & Walter Schaeken. 2007. When people are more logical under cognitive load: dual task impact on scalar implicature. *Experimental psychology* 54(2). 128–133.

What makes an inference robust?

- Dieuleveut, Anouk, Emmanuel Chemla & Benjamin Spector. 2019. Distinctions between primary and secondary scalar implicatures. *Journal of Memory and Language* 106. 150–171.
- Fox, Danny. 2007. Free choice and the theory of scalar implicatures. In Uli Sauerland & Penka Stateva (eds.), *Presupposition and implicature in compositional semantics*, 71–120. Palgrave.
- Fox, Danny. 2016. On why ignorance might not be part of literal meaning. Commentary on Marie-Christine Meyer at the MIT Workshop on Exhaustivity, Sept. 2016.
- Fox, Danny & Roni Katzir. 2011. On the characterization of alternatives. *Natural Language Semantics* 19(1). 87–107.
- Fox, John & Sanford Weisberg. 2019. *An R companion to applied regression*. Third. Thousand Oaks CA: Sage. <https://socialsciences.mcmaster.ca/jfox/Books/Companion/>.
- Franke, Michael. 2011. Quantity implicatures, exhaustive interpretation, and rational conversation. *Semantics and Pragmatics* 4(1). 1–82. <https://doi.org/10.3765/sp.4.1>.
- Gazdar, Gerald. 1979. *Pragmatics: implicature, presupposition, and logical form*. New York: Academic Press.
- Geurts, Bart. 2010. *Quantity implicatures*. Cambridge University Press.
- Goldstein, Simon. 2019. Free choice and homogeneity. *Semantics and Pragmatics* 12(23). 1–47. <https://doi.org/10.3765/sp>.
- Grice, Paul. 1975. Logic and conversation. in *The Logic of Grammar*, D. Davidson and G. Harman (eds), Encino, CA: Dickenson, 64-75.
- Groenendijk, J. A. G., T. M. V. Janssen & M. B. J. Stokhof. 1984. *Truth, interpretation and information : selected papers from the third amsterdam colloquium*. Dordrecht, Holland ; Cinnaminson, N.J.: Foris Publications. 182.
- Groenendijk, Jeroen & Martin Stokhof. 1984. *Studies on the semantics of questions and the pragmatics of answers*. Amsterdam: University of Amsterdam dissertation.
- Heim, Irene. 1994. Interrogative semantics and Karttunen's semantics for *know*. In Rhonna Buchalla & Anita Mittwoch (eds.), *Proceedings of iatl 1*. Hebrew University of Jerusalem.
- Hochstein, Lara, Alan Bale, Danny Fox & David Barner. 2016. Ignorance and Inference: Do Problems with Gricean Epistemic Reasoning Explain Children's Difficulty with Scalar Implicature? *Journal of Semantics* 33(1). 107–135. <https://doi.org/10.1093/jos/ffu015>. <https://doi.org/10.1093/jos/ffu015>.
- Jaeger, T Florian. 2008. Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language* 59(4). 434–446. <https://doi.org/10.1016/j.jml.2007.11.007>.

- Kamp, Hans. 1974. Free choice permission. *Proceedings of the Aristotelian Society* 74. 57–74.
- Kamp, Hans. 1978. Semantics versus pragmatics. In Franz Guenther & Siegfried Schmidt (eds.), *Formal semantics and pragmatics for natural languages*, 255–287. Dordrecht.
- Katzir, Roni. 2007. Structurally-defined alternatives. *Linguistic and Philosophy* 30(6). 669–690.
- Klinedinst, Nathan. 2007. *Plurality and possibility*. UCLA dissertation.
- Lenth, Russell V. 2021. *emmeans: Estimated Marginal Means, aka Least-Squares Means*. R package version 1.6.0. <https://CRAN.R-project.org/package=emmeans>.
- Marty, Paul. 2017. *Implicatures in the DP domain*. MIT dissertation.
- Marty, Paul & Jacopo Romoli. 2021. Presupposed free choice and the theory of scalar implicatures. *Linguistics and Philosophy*.
- Marty, Paul, Jacopo Romoli, Yasu Sudo, Bob van Tiel & Richard Breheny. 2020. Processing implicatures: a comparison between direct and indirect sis. *Talk presented at the first Experiments on Linguistic Meaning (ELM) Conference*. [https://campuspro-uploads.s3-us-west-2.amazonaws.com/2a1da9b9-a2cb-4a63-b175-3c919aa79aca/1fc946ff-35a6-427f-b473-efae316fc584/ELM-Directvs.IndirectSIs-Martyetal.\(2020\).pdf](https://campuspro-uploads.s3-us-west-2.amazonaws.com/2a1da9b9-a2cb-4a63-b175-3c919aa79aca/1fc946ff-35a6-427f-b473-efae316fc584/ELM-Directvs.IndirectSIs-Martyetal.(2020).pdf).
- Marty, Paul, Jacopo Romoli, Yasutada Sudo & Richard Breheny. 2021. Negative free choice. *Semantics and Pragmatics* 14. 13.
- Marty, Paul Pierre & Emmanuel Chemla. 2013. Scalar implicatures: working memory and a comparison with only. *Frontiers in psychology* 4. 403.
- Meyer, Marie-Christine. 2013. *Ignorance and grammar*. MIT dissertation.
- Noveck, Ira A. 2001. When children are more logical than adults: experimental investigations of scalar implicature. *Cognition* 78(2). 165–188.
- Pagliarini, Elena, Cory Bill, Lyn Tieu & Stephen Crain. 2018. On children’s variable success with scalar inferences: insights from disjunction in the scope of a universal quantifier. *Cognition* 178. 178–192. https://semanticsarchive.net/Archive/mE0YjdkN/PagliariniBillRomoliTieuCrain_DistributiveInferencesAcq.pdf.
- R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. <https://www.R-project.org/>.
- Ramotowska, Sonia, Paul Marty, Jacopo Romoli, Yasutada Sudo & Richard Breheny. 2022. Diversity with universality. In *Proceedings of the Amsterdam Colloquium 2022*. <https://www.dropbox.com/s/4vffehe2q2efljs/Proceedings2022-pages-258-264.pdf?dl=0>.
- Romoli, Jacopo. 2012. *Soft but strong: neg-raising, soft triggers, and exhaustification*. Harvard University dissertation. https://dash.harvard.edu/bitstream/handle/1/9909638/Romoli_gsas.harvard_0084L_10566.pdf?sequence=3.

- Romoli, Jacopo, Paolo Santorio & Eva Wittenberg. 2020. Alternatives in counterfactuals: what is right and what is not. Unpublished ms. University of Bergen, University of Maryland, and University of California, San Diego.
- Rothschild, Daniel & Stephen Yablo. 2018. Permissive updates. Ms, UCL and MIT.
- Santorio, Paolo & Jacopo Romoli. 2017. Probability and implicatures: a unified account of the scalar effects of disjunction under modals. *Semantics & Pragmatics* 10(13).
- Sauerland, Uli. 2004. Scalar implicatures in complex sentences. *Linguistics and Philosophy* 27(3). 367–391.
- Schulz, Katrin. 2019. The similarity approach strikes back: negation in counterfactuals. In *Proceedings of sinn und bedeutung* 22, vol. 2, 343–360. <https://ojs.ub.uni-konstanz.de/sub/index.php/sub/article/view/110>.
- Simons, Mandy. 2001. Disjunction and alternativeness. *Linguistics and Philosophy* 24. 597–619.
- Simons, Mandy. 2005. Dividing things up: the semantics of or and the modal/or interaction. *Natural Language Semantics* 13. 271–316. <https://doi.org/10.1007/s11050-004-2900-7>.
- Singh, Raj. 2008. On the interpretation of disjunction: asymmetric, incremental and eager for inconsistency. *Linguistics and Philosophy* 31. 245–260.
- Singh, Raj, Ken Wexler, Andrea Astle-Rahim, Deepthi Kamawar & Danny Fox. 2016. Children interpret disjunction as conjunction: consequences for theories of implicature and child development. *Natural Language Semantics* 24(4). 305–352.
- van Tiel, Bob, Paul Marty, Elizabeth Pankratz & Chao Sun. 2019. Scalar inferences and cognitive load. In *Proceedings of sinn und bedeutung*, vol. 23, 427–442.
- van Tiel, Bob, Elizabeth Pankratz & Chao Sun. 2019. Scales and scalarity: processing scalar inferences. *Journal of Memory and Language* 105. 93–107.
- Tieu, Lyn, Jacopo Romoli, Peng Zhou & Stephen Crain. 2016. Children’s knowledge of free choice inferences and scalar implicatures. *Journal of Semantics* 33(2). 269–298.
- Van Tiel, Bob & Walter Schaeken. 2017. Processing conversational implicatures: alternatives and counterfactual reasoning. *Cognitive Science* 105. 93–107.
- Willer, Malte. 2017. Widening free choice. In *Proceedings from the amsterdam colloquium 2017*.
- Zimmerman, Thomas Ede. 2000. Free choice disjunction and epistemic possibility. *Natural Language Semantics* 8(255–290).