Synthese (2021) 199:991–1018
https://doi.org/10.1007/s11229-020-02746-5

COMPUTATIONAL MODELLING IN PHILOSOPHY

Lying, more or less: a computer simulation study of graded lies and trust dynamics

Borut Trpin[1] · **Anna Dobrosovestnova**[2] · **Sebastian J. Götzendorfer**[3]

Received: 28 February 2019 / Accepted: 11 June 2020 / Published online: 30 June 2020

Abstract
Partial lying denotes the cases where we partially believe something to be false but nevertheless assert it with the intent to deceive the addressee. We investigate how the severity of partial lying may be determined and how partial lies can be classified. We also study how much epistemic damage an agent suffers depending on the level of trust that she invests in the liar and the severity of the lies she is told. Our analysis is based on the results from exploratory computer simulations of an arguably rational Bayesian agent who is trying to determine how biased a coin is while observing the coin tosses and listening to a (partial) liar's misleading predictions about the outcomes. Our results provide an interesting testable hypothesis at the intersection of epistemology and ethics, namely that in the longer term partial lies lead to more epistemic damage than outright lies.

Keywords Lying · Epistemic damage · Blameworthiness · Computer simulations · Formal epistemology · Ethics

> If falsehood had, like truth, but one face only, we should be upon better terms; for we should then take for certain the contrary to what the liar says: but the reverse of truth has a hundred thousand forms, and a field indefinite, without bound or limit (Montaigne).

✉ Borut Trpin
borut.trpin@lrz.uni-muenchen.de

Anna Dobrosovestnova
anna.dobrosovestnova@ofai.at

Sebastian J. Götzendorfer
sebastian.goetzendorfer@univie.ac.at

[1] Fakultät für Philosophie, Wissenschaftstheorie und Religionswissenschaft, Munich Center for Mathematical Philosophy, Ludwig-Maximilians-Universität München, Geschwister-Scholl-Platz 1, 80539 Munich, Germany

[2] Austrian Research Institute for Artificial Intelligence, Freyung 6/6, 1010 Vienna, Austria

[3] Department of Cognition, Emotion, and Methods in Psychology, Faculty of Psychology, University of Vienna, Liebiggasse 5, 1010 Vienna, Austria

1 Introduction

Lying is traditionally defined as stating that some proposition *A* is the case when *A* is believed to be false with the intention that the other person believes *A* to be true. The definition implies four necessary conditions for lying (see Mahon 2016):[1]

(C1) Making a statement
(C2) Believing that statement to be false
(C3) Addressing another person with that statement, and
(C4) Intending to deceive the addressee.

In this paper we will focus on (C2), the belief condition, because we consider it too strong of a requirement to classify a given act as lying. Most of the debates regarding the validity of (C2) have focused on whether a statement that is believed to be true (hence, not satisfying C2) may be classified as a lie if the remaining conditions are satisfied. Examples that are typically raised against (C2) include cases where the speaker believes a statement to be true but pragmatic implicatures suggest the contrary of what is said,[2] or the cases of double bluff[3] (Mahon 2016, 1.2).

Our objection to (C2) is different. We aim to show that (C2) is too strict and needs to be broadened. That is, the statements that are not believed to be outright false (hence, they fail to satisfy (C2)), but of which the speaker is *more certain* that they are false than true should also be considered as lies (we will call this C2* as a shorthand).[4] Specifically, we believe (1) that one may be lying without (outright) believing a statement to be false, (2) that this is very common, and (3) that these cases are, in the long run, epistemically more damaging than outright lies, where by epistemic damage we mean that one becomes more confident in false propositions.

By means of an illustration: Suppose Larry wants to go golfing tomorrow. He turns on the TV and listens to the weatherman's forecast. Now, suppose the weatherman knows there is a 30% chance of rain because this was the information he received from an advanced forecasting system. The weatherman could share the information with his audience through a full report by saying, e.g., "According to our models, there is a

[1] Although it could be said that either (C1) or (C3) is redundant, we stick to the traditional conditions of lying for the sake of the argument. The intuition behind (C3) is that one cannot lie if the addressee is not able to receive the lie. For instance, one cannot lie to a spy if the spy is (contrary to the speaker's suspicion) not monitoring the communication (see Mahon 2016, 1.3). We would like to thank an anonymous reviewer for pressing us on this point.

[2] For instance: Person A wonders whether B ate the whole cake. B, the culprit, states: "I ate a piece of it." but does not mention that he afterwards also ate the other pieces. B believes the statement to be true, yet it seems that he is lying because the statement implicates that he *only* ate one piece.

[3] For instance, in a game of poker, person A says: "I have a strong hand." B: "You're a liar—you want me to think you are bluffing and that you actually have a weak hand. But I will fold because you would not say this if you have not had a strong hand." In this case, A is accused of lying despite believing the statement ("I have a strong hand.") to be true.

[4] Note that one could say that (C2) already covers cases we aim at with (C2*) by arguing that to think some *p* likely suffices to believe it (see, e.g., Rothschild 2020). If we were to accept this claim, then (C2*) could be seen as a restatement and not a replacement of (C2). Be that as it may, this does not affect our investigation as our primary aim is not to redefine lying but rather to investigate consequences of lying when a statement is not outright believed to be false.

70%[5] chance that it will be sunny." or simply "It will *likely* be sunny." However, there seems nothing particularly wrong with him just stating that it will be sunny as this is very likely.

But could the weatherman lie in a case like this if he only unequivocally forecasts sunny or rainy weather?[6] Not in the sense covered by the traditional definition because he does not outright believe that it will either be sunny or not (C2). Nevertheless, suppose the weatherman wants to go golfing and wishes the golf course to be as empty as possible. It seems that he can at least partially lie by stating that it will rain (C1), although he is more certain that it will not than that it will (C2*). Conditions (C3) and (C4) are also satisfied in this case: the weatherman is addressing the audience in front of the TV screens (C3) and is claiming that it will be sunny with the intent to deceive his audience (C4) because he expects the viewers to change their plans, so that he can play golf on an empty course on a sunny day. To the best of his knowledge, there is a 30% chance that it will rain, but it is very likely that it will not and the audience will be deceived (and, subsequently, the golf course will be rather empty).

Although we are still in the realm of armchair philosophising, we predict that a vast majority of people would consider the weatherman to be lying (and not just deceiving) when he announces that it will rain.

In what follows, we investigate the different epistemic consequences imposed on dupes such as Larry when partial and categorical lies are compared.

2 Categorical and partial lies

We will distinguish between two senses of lying, which can be identified by the differences in the belief condition (C2/C2*):

1. Categorical lying: The speaker outright believes a statement to be false. (C2)
2. Partial lying: The speaker is more certain that a statement is false than true. (C2*)[7]

Partial lying affords liars greater deniability than categorical lying. As Marsili (2018) argues, partial lying provides a grey zone wherein liars strive to be because it is hard to distinguish between sincere and insincere assertions. We can easily see why this holds: in case of categorical lying insincere assertions turn out to be (mostly)[8] false, so a liar can be recognised as a liar and hence untrustworthy. On the other hand, in case of partial lying, insincere assertions often turn out to be false but they also often turn out to be true. This may give a dupe a sense of false security that the liar is simply not fully reliable. To return to our weatherman example: even if the weatherman constantly lies

[5] For the sake of simplicity, we assume that for both Larry and the weatherman rain and sun are the only possible and mutually exclusive forecasts. Hence, 30% chance of rain implies 70% chance of sun.

[6] The only statements of interest in this paper are outright statements. However, the weatherman could lie in the traditional sense if he reported another chance of rain/sun, e.g., "There is a 50% chance of sun tomorrow.", a statement he clearly does not believe. We would like to thank an anonymous reviewer for pointing this out.

[7] Note that (C2*)—in contrast to (C2)—overs both partial and categorical lying.

[8] Not all categorical lies are falsehoods because a liar could simply be wrong.

about the weather, his forecasts would still turn out to be to some degree correct. The weather is not fully predictable, hence even if one lies about it, the lie often turns out to be true. However, no-one, bar a hypothetical clairvoyant, may be fully reliable about uncertain events. Hence, when one is partially lying, they can easily argue that it is human to err. For instance, suppose Larry happens to see the weatherman golfing on a day when rain was predicted. The weatherman could argue that weather forecasts are not completely reliable. Combined with the fact that partial lies turn out to be true much more often than categorical lies, this means that it takes longer to assess how reliable a person is, which may—so we argue—lead to more epistemic damage than close encounters with "traditional" outright liars. In other words, although outright lies make a trustful dupe become more confident about false propositions (i.e., they make her suffer epistemic damage), they also lead to more directly decreased trust towards the liar. Partial lies are in this sense more damaging, and partial lying is more blameworthy, because the dupe retains trust towards the liar and confidence in false propositions in the longer run.

Before we proceed to our investigation, we will define more precisely what partial lying denotes in line with the recent debates in epistemology and ethics of partial lying. In particular, Marsili (2014) proposes a new definition of lying in which (C2) (the belief condition) from the traditional definition is replaced by the comparative insincerity condition (CIC):

S believes A more likely to be false than true. (CIC)

The condition allows us to extend the definition of lying to exactly the cases we consider to be partial lying, and which the previous definitions of lying omitted. It also provides the basis for a probabilistic interpretation, which we are to follow as it allows us to formalise and computationally model and simulate partial lying. Our understanding of (C2*) may thus be formulated in the following way. For any A:

$$0 \leq \Pr(A) < 0.5 \qquad (\mathrm{C2_{Pr}})$$

where $\Pr(\cdot)$ is the liar's probability function and A is the proposition that the liar states in order to deceive the addressee.[9] This means that on our understanding of lying, a person that is more confident that some proposition A is false than true is lying when she asserts that A is the case to another person. When condition ($\mathrm{C2_{Pr}}$*) is satisfied in addition to the other conditions (C1, C3, C4), we say that a person is partially lying:

$$0 < \Pr(A) < 0.5 \qquad (\mathrm{C2_{Pr}}^*)$$

In other words, a person may only be partially lying when she is not fully certain about the proposition in question.[10]

[9] There is an ongoing debate whether (C4), the intention to deceive, is necessary for lying (e.g. Sorensen 2007; Marsili 2014). We retain the condition because we do not have any reasons to criticise it in scope of this paper. However, this should not indicate that we endorse (C4).

[10] The belief condition (C2) and (CIC) as a candidate for its replacement both operate with beliefs. The relationship between beliefs and degrees of belief is complicated, but it is not our focus. Instead, we simply

However, according to Krauss (2017), this condition in both its informal (CIC) and the probabilistic version ($C2_{Pr}$) fails to cover some problematic cases in which we blame the speaker for lying. To illustrate: suppose the weatherman is talking to Larry. The weatherman's degree of belief in rain is 0.51, but they both know Larry is agnostic with respect to rain (i.e., his degree of belief in rain is 0.5). When the weatherman asserts that it will in fact rain, he can—assuming Larry trusts him—expect Larry to move his degree of belief in rain from 0.5 to, for instance, 0.8, yet his own degree of belief will remain nearly agnostic (0.51). The author argues that we blame the weatherman in a case like this for the same reasons we blame liars, yet (C2*) fails to recognise such cases as lying, simply because the weatherman is more confident that it will rain than that it will not. Krauss (2017) then proposes a replacement for (C2), the worse-off requirement (WOR):

> The expected epistemic damage to the audience, with respect to p, by the speaker's lights, conditional on the audience trusting her with respect to p, at all, is greater than 0. (WOR)

In other words, lying—be it partial or categorical in the sense described above—hinges on the expected epistemic damage brought upon the audience. Expected epistemic damage here denotes that the speaker *expects* a trustful dupe to update her credence in the asserted proposition, p, in such a way that it ends up further from what the speaker considers to be appropriate.[11] We agree with Krauss (2017) that a person who is only slightly more certain that A is the case than $\neg A$ but outright states A is blameworthy for similar reasons we blame liars if that person expects the addressee to incur epistemic damage as a consequence. That is, the person stating A is blameworthy because the addressee is expected to end up in a worse epistemic state, a state that the speaker considers to be further from the truth. In other words, the speaker is leading the dupe astray from the truth as known by the speaker. Yet this should not be considered an act of lying but rather intentional misleading, or partial truth-telling if the speaker does not intend to deceive the audience. Even if our concern (lying vs. "merely" misleading vs. partial truth-telling) may be seen as splitting hairs, the definition of lying that stems from Krauss (2017)'s revision of (C2)—which is (WOR)—also gives rise to its own host of problems as it provides neither necessary nor sufficient conditions of what we consider to be (partial) lying. For instance, suppose A is fully certain that p and B is fully certain that $\neg p$, and A knows this. When A asserts "$\neg p$", she does not expect B to be worse off as B is already in the worst epistemic position (with respect to p, according to A), yet the statement is clearly a lie which is expected to keep the dupe

Footnote 10 continued
separate partial and categorical lying at the line of full and less than full certainty and leave a more detailed investigation of the relationship between full and partial beliefs—in the thematic scope of partial lying—for future research.

[11] For instance, the speaker is blameworthy if he expects the audience to update the credence in p from 0.5 to 0.8 when he considers the appropriate credence in p to be 0.55, but not when he considers it to be 0.7. In the first case the audience's "distance" from what the speaker considers to be the appropriate credence in p (0.55) is expected to increase from 0.05 to 0.25 as a consequence of the statement. In the latter case the "distance" from the appropriate credence in p (0.7) is expected to decrease from 0.2 to 0.1. The statement is therefore expected to be epistemically damaging to the audience (from the speaker's perspective) in the first and beneficial in the latter case.

in the wrong. (WOR) is therefore not a necessary condition for lying. Similarly, when both A and B are 0.8 certain that p and A knows that, but still asserts "p" expecting B to become more confident about p, (WOR) identifies this case as a lie because A may expect B to end up further from what A considers to be the appropriate credence in p. (WOR) is therefore also not sufficient to distinguish between what we consider lying and what we consider to be an act of trying to tell the truth (both counterexamples are due to Benton 2018). We will therefore stick to our revised belief condition (C2*) and its probabilistic counterpart ($C2_{Pr}$).

Nevertheless, we will show below that our results also confirm some of the worries posed by Krauss (2017). When a categorical statement is made in uncertain situations, it will lead to foreseeable epistemic damage by decreasing epistemic accuracy of the addressee even if the speaker is more certain about the truth of the statement than its falsehood. Yet many, perhaps most, statements we make in everyday conversations are categorical ("This is so-and-so") as we typically do not describe our confidence, for instance, folk-probabilistically ("I would bet x units that this is so-and-so"), comparatively ("I believe that this is like this more than I believe that it is like that") or in some other way. These statements where the speaker is trying to tell the truth (partial truth-telling) are therefore also important for our research.

The condition (WOR) is also important for another reason: it includes trust. Trust is typically only briefly, if at all, addressed in discussions of lying, although it is central for analysing the consequences of lying in the longer term. To continue with our previous illustration: Suppose the weatherman is fairly confident it will be sunny but states that it will rain, so that he can go golfing on an empty course. Larry trusts him and changes his plans to go golfing. However, he sees that it is sunny the next day and decides to pay a visit to the golf course. He notices that the weatherman himself is golfing, becomes suspicious and stops trusting him. Hence, the next time the weatherman forecasts rain, Larry interprets the forecast to mean just the opposite (concluding that rain actually means sun because the weatherman is a liar): he therefore goes to the golf course despite (or because) of the forecast.[12] The concept of trust is essential for an analysis of lying because if a person S does not trust the person P, then S may simply ignore P or even believe the very opposite of what P says.

We believe that trust, like belief, is a graded concept.[13] We may trust, distrust, neither trust nor distrust someone with respect to something, or we may be leaning towards one of the extremes. We can quantitatively describe trust as a probabilistic concept which can be measured by τ, $\tau \in [0, 1]$, with $\tau = 0$ describing categorical distrust, $\tau = 0.5$ neither trust nor distrust and $\tau = 1$ categorical trust.[14]

Assuming we want to minimise epistemic damage, it is much easier to determine how much trust we should put towards a person that is categorically lying (or cate-

[12] The weatherman examples are based on the plot of an episode from the series Curb Your Enthusiasm where the main character, Larry, eventually ends up playing golf in heavy rain.

[13] Although our conception of trust as a graded concept does not follow any established conception, trust has been considered as a graded concept before (see, e.g., Lorini and Demolombe 2008). In addition, there are other related analyses of trust in a Bayesian framework that have been successfully applied to various phenomena (e.g., the problem of overconfidence; Vallinder and Olsson 2014).

[14] As we show below, the range of τ in our model is (0, 1) due to the specific formula for updating trust that we use (LSR).

gorically telling the truth) than in cases of partial lying (or partial truth-telling in the sense of the above-mentioned examples by Krauss 2017), especially if the pattern is repetitive. That is, if A constantly categorically lies, then the appropriate response seems to be to distrust A and assume that when that person asserts B, they actually believe $\neg B$. Such patterns of constant categorical lying, however, are non-realistic, and hence the problem of the appropriate level of trust (in the described sense) towards a source arises. How much should we trust a person that is either not constantly lying or when she is partially lying?

The question that we set out to analyse was how much difference there is between the categorical and partial lying when we assess the epistemic damage while accounting for the dynamically evolving degree of trust towards the source. Contrary to Krauss (2017) who refers to *expected* epistemic damage, we wanted to assess *actual* epistemic damage suffered by a dupe. The former notion assesses how much more confident a dupe is expected to become about false propositions if she trusts a liar. Its focus is therefore on the liar's blameworthiness (how much damage a liar is expecting to impose on a dupe). Actual epistemic damage, on the other hand, focuses on the dupe's actual situation, that is, on how much more confident about false propositions the dupe actually becomes. This notion is more relevant for us as our focus is on the consequences of being lied to from the dupe's perspective.

We were also interested in answering whether partial lying can be further analysed in a more fine-grained fashion and, again, how partial lying of different degrees affects the epistemic situation of a dupe. To explore these questions we developed a computer simulation based on an agent (the dupe) who updates her beliefs based on lies of a varying severity: from categorical lies to statements where the speaker is agnostic with respect to the truth of what is said (as such, these statements are not proper lies but closer to bullshit in the sense of Frankfurt 2005). We also incorporated the dupe's dynamic degree of trust in the source. The model of belief updating is in line with the principles of Bayesian epistemology (see, e.g., Hartmann and Sprenger 2010), so the dupe can be described as a rational agent. This in turn allows us to precisely investigate the consequences of being lied to from the perspective of an arguably ideally rational agent.[15]

3 Belief updating in (un)trustworthy circumstances

But how exactly could we investigate the consequences of being lied to in all these different senses that we described above? This brings us to a scenario where we can

[15] There is a possibility that the evolution and the importance of not being deceived could lead to the development of mechanisms that help us, non-ideal human agents, to defend ourselves against the consequences of being lied to in different ways than idealised Bayesian agents in our model (for an example of a similar "cheater detection" mechanism, see Cosmides and Tooby 1992). However, using idealised Bayesian agents allows us to investigate the problem more precisely and efficiently. We believe our findings could be applied to non-ideal agents too, although we consider our present findings more as an exploratory step that could provide the basis for further empirical research.

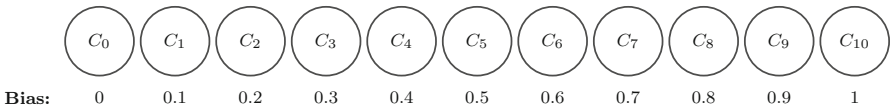

Fig. 1 Coins that the magician may flip and their bias to land heads

easily determine to what degree a person believes what she is stating and, consequently, how severe of a lie the statement is: a game of determining how biased a coin is.[16]

Suppose there are two persons, a magician and a player. The magician shows the player 11 coins with varying biases to land heads. Specifically, the coins are biased to land heads from 0 to 1 in 0.1 increments but otherwise identical and indistinguishable. This means that Coin 0 (C_0) never lands heads, C_1 on average lands heads 10% of the time, and so on. C_5 is a fair coin and C_{10} is fully biased towards heads (i.e., always lands heads); see also Fig. 1. The player believes that the coins indeed have the described biases, for instance because she has the opportunity to flip each coin as long as she wants to confirm that in the long run they indeed land heads according to the bias. For simplicity's sake also suppose that the coins may only land heads or tails. That is, the coins cannot land on the side and both the magician and the player acknowledge this. Finally, the game proceeds in the following manner: the magician selects one of the 11 coins and starts repeatedly flipping it in front of the player. The player then needs to determine which of the coins the magician selected.

Because the magician knows which of the coins is used in the game, it is straightforward to see how lying to varying degrees may be incorporated into the game. Suppose the magician does not only flip the coins but for some undisclosed reason also provides the player with tips regarding what side the coin will land on. For instance, if the magician selects a coin that is fully biased toward heads (C_{10}), then stating that the coin will land tails represents a categorical lie. The magician believes (with full certainty) that it will land heads, but states what he believes is false (that it will land tails). Similarly, the magician could be partially lying if the coin in question is, say, C_9 and hence 90% biased toward heads. If the magician says that this coin will land tails, the condition ($C2_{Pr}*$) for partial lying is satisfied: the magician is more certain that the coin will land heads than tails, yet he asserts that it will land tails. Importantly, this coin may actually land tails, in which case the statement would not be false.

This also allows us to assess the severity of lying: when the magician says the coin will land tails knowing that C_{10} is selected, the lie is as severe as possible—the statement is necessarily false. The severity of the lie decreases if the same statement is used when C_9 is used in the game, and further decreases when C_8, C_7, or C_6 is used. Finally, saying the coin will land tails when C_5 (a fair coin) is used, is more akin to bullshit in Frankfurt (2005)'s sense as the magician has no inclination toward the truth or falsity of the statement. Although stating that the coin will land tails always imposes epistemic danger on a trustful player unless C_0 (always tails) is used in the game, we will describe such statements when C_1, C_2, C_3 or C_4 is used, as partial truth-telling (in analogy with partial lying). Note that we refer to the detection of biased coins

[16] Using a game of determining the bias of a coin is common in debates regarding epistemic performance (see, e.g., Douven 2013; Trpin and Pellert 2019).

Table 1 The evolution of trust in a potential scenario

Round	0	1	2	3	4	5	6	7	8	9	10
Statement		"T"	"T"	"T"	"T"	"T"	"T"	"T"	"T"	"T"	"T"
Toss		T	T	T	H	H	H	H	H	H	H
Trust	0.5	0.66	0.75	0.8	0.66	0.57	0.5	0.44	0.4	0.36	0.33

throughout the paper because this makes the modelling more straight-forward, but we could also assess other examples of partial lying in a similar manner. For instance, if the weatherman from the introductory example knows the probability of sun is 90%, yet says that it will rain, then the situation is analogous to that of flipping C_9 (coin with a 90% bias for heads) and saying it will land tails. The insights from the game of detecting the bias of a coin might therefore be applied to more realistic scenarios by analogy.

What can we assume will happen in such a game? By means of an illustration, suppose the magician selects C_7, the coin that is 70% biased toward heads. Suppose also that the magician always says that the coin will land tails and the sequence of the coin landings is T, T, T, H, H, H, H, H, H, H, where T stands for tails and H for heads. Obviously, the magician could only truthfully say that the coin will land on a specific side if the coin is fully biased (i.e., C_0 or C_{10}), but let us assume that he claims that he has a somewhat clairvoyant intuition about these things. The player ought to be sceptical about such claims, but when the coin lands tails three times in a row just like the magician predicted, the player might trust him and take the tips as additional pieces of information that may help her in her belief updating process. Yet when the prediction finally turns out to be false, she may either distrust him completely and ignore him or, on the other hand, decrease her level of trust towards him with respect to coin toss predictions. But how exactly should she calibrate her level of trust?

Because the player initially has no information about the reliability of the source and because we want to avoid the problem of assigning zero probability to hypotheses that are potentially true (this problem is addressed in more detail at the end of this section), we use a general form of Laplace succession rule (LSR) as a way of determining player's trust. The level of trust is therefore calculated by a simple formula (for a derivation see, e.g., Robert 2007, pp. 180–181):

$$\tau = \Pr(E|x_1) = \frac{x_1 + 1}{x + 2} \qquad \text{(LSR)}$$

where E stands for the proposition that the magician's latest prediction is true, x_1 for the number of past true predictions and x for the number of all predictions. By means of an example, this results in the development of trust towards the source[17] with respect to the coin toss predictions as presented in Table 1.

But how exactly should these levels of trust be taken into account when updating beliefs? That is, how could this additional level of information be useful for the player? Recall that the player in our examples is a Bayesian agent and updates her beliefs by

[17] In our examples, the source stands for the magician because he is the source of the statements.

Bayesian conditionalisation. In other words, when the player observes that the coin landed heads, she updates her beliefs in the following manner. For any $i \in L$, $L = \{i | i \in \mathbb{N}_0, 0 \leq i \leq 10\}$:

$$\Pr^*(H_i) = \Pr(H_i|Heads) = \frac{\Pr(Heads|H_i)\Pr(H_i)}{\sum_{j=0}^{10} \Pr(Heads|H_j)\Pr(H_j)} \tag{BC}$$

where $\Pr(\cdot)$ stands for the prior and $\Pr^*(\cdot)$ for the posterior probability function, and H_i for any of the 11 hypotheses about the coin that is used in the game (e.g., H_4 denotes the hypothesis that C_4 is used; see also Fig. 1). If the coin lands tails, we simply substitute $Heads$ with $Tails$. The likelihoods are also well-defined, i.e. for any $i \in L$, $\Pr(Heads|H_i) = i/10$, and $1 - i/10$ for tails.

The situation is different, however, before the player observes the outcome of the coin toss and is merely tipped off about it by the magician. That is, for any $i \in L$, the player updates in the following way:

$$\Pr^*(H_i) = \Pr(H_i|\text{`}Heads\text{'}) = \frac{\Pr(\text{`}Heads\text{'}|H_i)\Pr(H_i)}{\sum_{j=0}^{10} \Pr(\text{`}Heads\text{'}|H_j)\Pr(H_j)} \tag{BC$_\tau$}$$

where '$Heads$' represents the statement that the coin is predicted to land heads. Again, the formula also holds when the player is told the coin will land tails (by substituting '$Heads$' with '$Tails$').

The two belief updating formulas, (BC) and (BC$_\tau$), at the first glance look very similar—as expected as they are both instances of Bayesian conditionalisation—but they are importantly different. Looking at the right-most fractions, we notice that belief updating hinges on prior probabilities for each of the coins and the likelihood of the coin either landing heads or tails (BC), or being told that the coin would land heads or tails (BC$_\tau$). The former likelihood is well-defined, as we mentioned.

But what is the likelihood that the magician would state that the coin will land heads given that a specific coin is used in the game? That is, how do we determine the value of $\Pr(\text{`}Heads\text{'}|H_i)$? This is where the notion of graded trust comes into play. Recall that we observed that a completely untrustworthy source is expected to say just the opposite of what the case is and a trustworthy source is expected to state what actually is the case. The likelihoods $\Pr(\text{`}Heads\text{'}|H_i)$ that the player will therefore assign in a given round of the game will depend on her degree of trust, τ, towards the source. More specifically, the likelihoods may be defined in the following way for any $i \in L$:

$$\Pr(\text{`}Heads\text{'}|H_i) := \tau\Pr(Heads|H_i) + (1 - \tau)\Pr(Tails|H_i) \tag{Lkh$_\tau$}$$

That is, if $\tau \approx 1$, the likelihoods approximate objective likelihoods. If $\tau \approx 0$, they are just the opposite.[18] The other possibilities fall in-between (see Table 2 for some examples).

Defining the likelihoods in this way comes with some advantages: if the source is estimated to be neither trustworthy nor untrustworthy (i.e., $\tau = 0.5$), then the likeli-

[18] Similarly as the likelihoods for the updating on observations, $\Pr(\text{`}Tails\text{'}|H_i) = 1 - \Pr(\text{`}Heads\text{'}|H_i)$.

Table 2 Select likelihoods for the statement "Heads" given each of the coins and various levels of trust

$\Pr(\text{`}Heads\text{'}\mid H_i)$	H_0	H_1	H_2	H_3	H_4	H_5	H_6	H_7	H_8	H_9	H_{10}
$\tau = 1$	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
$\tau = 0.75$	0.25	0.3	0.35	0.4	0.45	0.5	0.55	0.6	0.65	0.7	0.75
$\tau = 0.5$	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
$\tau = 0$	1	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0

Note that it is a consequence of Laplace succession rule that τ can converge toward but it can never be equal to 1 or 0. We include $\tau = 1$ and $\tau = 0$ in Table 2 to illustrate what likelihoods would approximate if τ was very close to its lower or upper bound

hood for the statement is the same for all hypotheses (0.5), and hence no updating takes place. Similarly, if the source is almost completely trustworthy ($\tau \approx 1$), the updating approximates standard Bayesian conditionalisation on observations. However, if it is non-extreme, then the updating that follows from such likelihoods is more conservative in the sense that the likelihoods are relatively scaled down and the probabilities for each hypothesis consequently change less relative to their prior probabilities.

Two observations are to be made at this point: first, it follows from Laplace succession rule and our definition of the likelihood $\Pr(\text{`}Heads\text{'}\mid H_i)$ that if the player cannot decide whether the magician is (un)trustworthy (e.g., at the beginning of the game or if the predictions are correct exactly 50% of the time), then no belief updating takes place, i.e., the player is "ignoring" the source (this is always the case when $\tau = 0.5$ because the likelihood $\Pr(\text{`}Heads\text{'}\mid H_i)$ is the same for all i). This seems reasonable because she cannot make any use of these predictions.

Second, an advantage of the way we model the dynamics of trust is that τ at most approaches extreme probabilities (0 or 1) but never reaches these values. The importance of this consequence can be demonstrated with an example. Suppose the coin is fully biased towards heads and the magician truthfully predicts: "Heads", "Heads", "Heads", …, 1,000,000 times in a row. It intuitively seems that in such a case the level of trust towards the magician, τ, could be 1. However, suppose the magician decides to lie and predicts the coin would land tails in the next round. If the player considers the magician to be a fully trustworthy clairvoyant and fully trusts him ($\tau = 1$), she takes the magician's prediction to be as relevant as her actual observation of the coin landing heads (see line 1 in Table 2). This leads to a big problem because if the coin were to land tails, then (BC)—to which updating reduces when $\tau = 1$—would lead her to update the probability of the coin being fully biased to 0 ($\Pr(H_{10} = 0)$)—if a coin lands tails, it cannot be fully biased towards heads. Because hypotheses with zero probability cannot be resurrected in Bayesian epistemology, this means that the player could thus never identify the coin that is used in the game. Laplace succession rule as a way of updating trust successfully avoids problems of this sort because τ is always greater than 0 and less than 1.[19]

[19] We are very thankful to an anonymous reviewer for the suggestion to use LRS as it helps us avoid the problems of another trust updating function we used in an earlier version of this paper.

4 Computer simulations of belief updating in (un)trustworthy circumstances

A limitation of the previous philosophical debates about lying, and particularly about partial lying in comparison to categorical lying, is that they remain on a relatively abstract level and sometimes rely on unpersuasive examples. There is nothing wrong with abstract analyses like this; on the contrary, it is how philosophy usually progresses. However, a question which may be raised regarding the phenomenon of interest is what happens in the longer term, i.e., when the audience changes the level of trust towards the liar with respect to his statements? It is possible to reason about this question from the armchair: one could, for instance, speculate that a more cautious audience—that is, the audience with lower degrees of trust toward the speaker—may end up in a better position because they would not take the statements for granted. Similarly, one could reason about the different epistemic consequences given varying severity of lies. For instance, one may speculate that categorical lying may easily be spotted, while on the contrary, severe but not categorical partial lies (e.g., when the speaker is, say, 90% certain that the statement is false) may present the hardest problem for the audience because the statements would often mislead them, yet they would also be true in many cases and the speaker may be able to deny responsibility because the statements were about an uncertain event.

To tackle these issues more efficiently and come to more definitive resolutions we developed a computational model.[20] Computational models offer distinct advantages when investigating systems too complex to deal with directly when it comes to clarity and completeness of the investigation, better options of exploration as well as comprehensibility of findings (Fum et al. 2007), characteristics we wanted to take advantage of within the realms of this work.

On the basis of our computational model we simulated the following cases. For the simulations we first generated sequences of 500 coin tosses for each of the 11 possible coins. We settled on 500 rounds because, in cases like ours where the agent is operating with only 11 hypotheses, 500 rounds in general suffice for convergence of the probability of one of the hypotheses towards full certainty and the others towards 0. After that we generated a list of statements in accordance to various understandings of partial lying. Particularly, we generated three lists of lies by the following principles:

1. Simple lying: State what is the least likely outcome of a coin flip[21]
2. Gambler's lying: Flip the coin in secret and state the opposite of the outcome, then flip it again in front of the player
3. Clairvoyant lying: Knowing exactly what the outcome will be, state the opposite[22]

We then generated the lists of actual statements that were to be used in the game. To make the statements more realistic (and more complicated for the player), we introduced varying inclinations towards lying from the magician's perspective. The inclinations were probabilistic in nature: if the magician is, say, 60% inclined towards

[20] Source code is available on request from the corresponding author.

[21] For simplicity: when the coin is fair, the magician always states that it will land tails.

[22] This type of lying always represents categorical lying because the clairvoyant knows (and, hence, believes and is certain of) the outcome of the coin toss.

Table 3 An example of five "tips" (statements) when a coin is 0.7 biased toward heads and the magician is 0.6 inclined towards lying

	Actual outcome:										
	H	H	H	T	H						
Simple lying:	T	T	T	T	T	**Actual statement:**	H	T	H	T	T
Gambler's lying:	H	T	T	H	T	**Actual statement:**	T	H	T	H	T
Clairvoyant lying:	T	T	T	H	T	**Actual statement:**	H	T	T	T	T

lying, we take each of the three lists of lies and revise them in such a way that each of the statements has 60% probability to be a lie (in the sense of lying as described with the above three principles). His "tips" for the player could thus be similar to those described in Table 3. We generated lists with 11 different inclinations toward lying (from 0 to 1 with 0.1 increments).

Note that when the inclination to lie is 0, we actually simulate truthful statements, but these statements may nevertheless be deceiving, which is very close to the problems Krauss (2017) poses against (C2*) as a replacement of (C2) in the definition of partial lying. For instance, if the magician is flipping a coin that is, e.g., 0.8 biased toward heads, then he would always state that it is going to land tails if his lying style was that of simple lying. Yet, if the magician's inclination to lie was 0, he would always state that the coin is to land heads in the first style (simple lying) and a mixture of mostly true but also false predictions in the second style (gambler's lying). He would only be able to make correct predictions if he was a truthful clairvoyant. Hence, a trustful player would epistemically be worse off in any case unless the magician was a clairvoyant or the coin was fully biased. As noted, we consider these cases not to be cases of lying but merely cases of either misleading or partial truth-telling. Partial truth-telling is not our focus in the present paper, but it is worth pointing out that in the longer-term we obtain results that are symmetrical to the partial lying. That is, if a source asserts some statement while being, e.g., 0.8 certain that the statement is false or 0.8 certain that it is true, the epistemic consequences for the dupe are almost identical: the dupe is worse-off to the almost same degree in both scenarios. Partial truth-telling can deceive the addressee in the same way as partial lying. We explain this symmetry in a bit more detail at the end of the next section where we discuss the results of our simulations.

To return to our simulations: once we had the lists of actual outcomes and the statements (in accordance with the lying style and the inclination towards lying), the simulations proceeded in a straight-forward fashion. At the beginning of the standard scenario simulation[23] the agent starts with $\tau = 0.5$ (neither trusting nor distrusting the source in the first round)[24] and all hypotheses about the coins are equiprobable (because there is no reason any hypothesis would initially be favoured). The agent is then introduced to the magician's statement (in 11×3 variations according to different inclinations to lying and lying styles), updates her beliefs in line with the Bayesian

[23] We discuss a variation of this scenario in line with the truth bias, i.e. where the dupe originally blindly trusts the magician, in Sect. 7.

[24] This is because of Laplace succession rule, $\tau = (t + 1)/(t + f + 2)$, where t represents the number of true and f false predictions. When $t = f = 0$, $\tau = 1/2$.

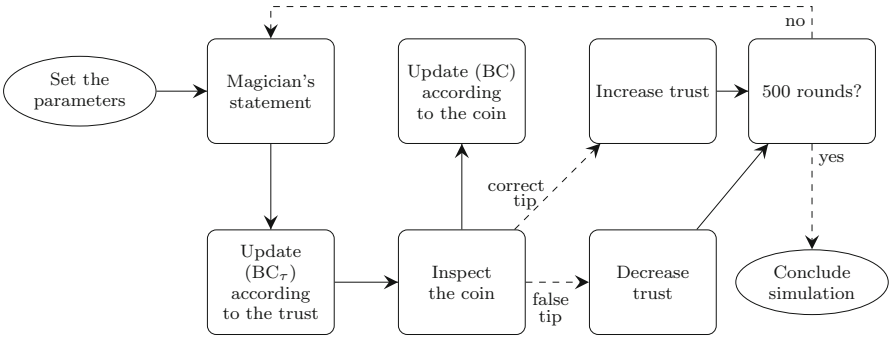

Fig. 2 Flowchart of the computer simulations (dashed lines are conditional)

formula we introduced above (BC_τ), then observes the actual outcome of the coin toss and updates her trust towards the source and updates her beliefs again (by simple Bayesian conditionalisation; BC) as she was introduced to new information (the actual outcome of the coin toss). This cycle then repeats until 500 rounds have passed, which concludes one simulation. However, because we were interested in statistically reliable results, we repeated each simulation (in various settings) 1000 times and then looked at average values for each of the steps.

These simulations were then run for each of the 11 possible coins, so that we could investigate the effect of (partial) lying on the epistemic state of the player. This resulted in $3 \times 11 \times 1000 \times 11 = 363,000$ simulations (lying styles $\times$ inclinations to lie $\times$ repetitions $\times$ coin biases).[25] As a control, we used the same sequences of coin tosses for each of the biases and simulated belief updating when the "tips" are ignored throughout the game. The procedure of the simulations is described schematically in Fig. 2.

5 Epistemic consequences of being (partially) lied to

Considering that our goal is to provide an insight into epistemic consequences of being (partially) lied to, we also need to have a measure with which we can compare what happens in various settings. To achieve this, we used two measures: Brier scores and a simple measure of how misguided the player's expectation of the outcome of the coin toss was compared to the coin's actual bias.

Brier scores represent a proper score function that measures the (in)accuracy of probability distributions.[26] It originates from meteorology (Brier 1950) where it was

[25] As observed by an anonymous reviewer, only 6 coin biases are relevant, i.e. either those that are more likely to land heads or those that are more likely to land tails as the severity of lying is symmetric, only the content is exchanged ("Heads" instead of "Tails"). The number of simulation could thus be substantially reduced. Nevertheless, this is also the reason why we only focus on one-sided coin bias throughout the paper.

[26] It should be noted that there are also other scoring rules that are frequently referred to in debates regarding epistemic accuracy. For instance, Levinstein (2012) argues that quadratic scoring rules, which also include Brier score (BS), should be rejected in favour of other rules, for instance the logarithmic rule:

originally used to assess the accuracy of probabilistic weather forecasts. However, it is also commonly used in debates that investigate epistemic accuracy in formal epistemology (e.g. Leitgeb and Pettigrew 2010a, b). The scoring rule is provided by the following formula:

$$BS = (1 - \Pr(H_t))^2 + \sum_{i=1}^{10} \big(0 - \Pr(H_f)\big)^2 \qquad \text{(BS)}$$

where H_t represents the true hypothesis (the coin that is actually used in the simulation) and H_f represents the other (false) hypotheses. As we can see from the formula, it means that the higher the Brier score is, the higher is the epistemic damage caused by the liar (this is also why the score is sometimes called Brier penalty).

By means of an example, if the coin used in the game is C_5 (fair coin), then the true hypothesis is H_5 (see Fig. 1 and Table 2). If the player is fully certain that H_5 is true, i.e., if $\Pr(H_5) = \Pr(H_t) = 1$, the Brier score returns the minimal possible value, 0, because in this case $BS = (1 - 1)^2 + 10 \times (0 - 0)^2 = 0$. On the contrary, if the player is mistakenly fully certain that, say, C_4 is used in the game, the Brier score will reach the maximum possible value, 2. This is because the true hypothesis, H_5, is in this case assigned zero probability[27], so $BS = (1 - 0)^2 + 9 \times 0 + (0 - 1)^2 = 2$. Brier scores close to 0 therefore indicate that the player has an accurate subjective probability distribution while those close to 2 indicate high inaccuracy. We computed the score after every round of the game (i.e., after updating in line with the magician's statement and observing the actual outcome of the coin toss).

However, the score (BS) is in some sense biased toward true hypotheses, so it does not always provide us all the information we are interested in. Suppose the coin that is used in the game is 0.9 biased toward heads. Now consider two different outcomes: one in which the player is 0.9 certain that the coin with 0.7 bias is used (H_7) and—for simplicity—0.01 certain that each of the other 10 coins is, and another in which the player is 0.9 certain that the coin with 0.8 bias is used (H_8) and 0.01 about the others. Brier score will be the same in both cases because the probability for the true hypothesis is the same (0.01): $0.99^2 + 0.9^2 + 9 \times 0.01^2 = 1.791$. The score thus indicates a highly inaccurate probability distribution in both cases.

Yet there seems to be an intuitive difference between the two outcomes: it seems that the player in the first outcome (high probability for H_7) is in some sense further away from the truth. This is because the expectation of the next outcome (i.e., the

Footnote 26 continued
$LR = -log(\Pr(H_t)$, where H_t stands for the true hypothesis. In our cases both BS and LR return similar results as they both measure inaccuracy with respect to the probability of the true hypothesis. However, the range of LR is $[0, \infty)$ while that of BS is $[0, 2]$. The means calculated by LR may therefore be distorted: if the true hypothesis is assigned very low probability in even a single simulation, its LR will be very large and the mean will be unrepresentative. A more detailed discussion of various scoring rules exceeds the goals of this paper, but it is also not necessary as BS arguably provides an acceptable measure of (in)accuracy (see, e.g., Leitgeb and Pettigrew 2010a, for a detailed discussion) and because in our cases both BS and the other major contender, LR, provide comparable insights. We would like to thank an anonymous reviewer for making us clarify this point.

[27] Mutually exclusive and jointly exhaustive hypotheses have to sum up to 1, i.e., $\sum_{i=0}^{10}(H_i) = 1$. Hence if $\Pr(H_4) = 1$, all other hypotheses (including the true H_5) are assigned zero probability.

probability of the coin landing heads) is further from the actual bias of the coin. In the first case, the expectation of heads is $\Pr(Heads) = 0.678$, while in the second case it is $\Pr(Heads) = 0.767$.[28] Ideally, the expectation of heads would be 0.9 because this is the objective probability of the coin to land heads, but nevertheless, although both outcomes obtain the same Brier score, the outcome in the second case is in this sense "closer to the truth", a fact that Brier score fails to detect. We, therefore, calculated both Brier scores and the absolute difference between the expectation of the outcome and its objective probability.[29] These two scores were then used to measure the epistemic damage incurred by being lied to and to compare the different outcomes in various settings. In other words, we considered a situation to be epistemically more damaging if the player's Brier scores were higher and if her expectation of the side the coin would land on was further away from the coin's actual bias.

Our simulations were conducted in many different variations as described above,[30] but we will mainly focus on the main insights of the large number of simulation studies which we ran. First, looking at the overall epistemic accuracy, i.e., the average Brier scores incurred in simulations, we find an interesting result: it is not the categorical lies that are the most epistemically damaging (given the specifics of our model). Neither are the bullshit statements or nearly bullshit statements (in the sense of Frankfurt 2005) the most damaging. Instead, the in-between partial lies turn out to be the most damaging. What we mean by this is that when a liar—regardless of the lying style, but especially in the cases of what we call simple lying—is constantly lying about something that is certainly false, in this case stating that a coin that always lands heads will land tails, a rational agent updating her beliefs and trust in the described ways will quickly realise that the source is untrustworthy. Hence, the statements will be interpreted as almost the exact opposite of what is said and she will not be misled. Similarly, when the source is making statements while being highly uncertain about either the truth or the falsehood of the statement in question—in our case, when the coin is fair or only slightly biased, then the statements made by the source are soon recognised as rather useless ($\tau \approx 0.5$). This results in the rational agent (almost) ignoring the source and does not lead to much epistemic damage because the agent (almost) only considers the actual observations when $\tau \approx 0.5$ (see row 3 in Table 2).

However, it is the cases where the statements by the source are mostly false but also sometimes true, hence, where the source is considered to be somewhat but not completely untrustworthy, that turn out to be the most epistemically damaging. The agent tries to use the evaluation of the source's untrustworthiness to their advantage, but ends up in a worse epistemic state than if the source was simply ignored. In our scenario this happens when the source is lying about the coins that are somewhat

[28] Both numbers are calculated by the law of total probability: $\Pr(Heads) = \sum_{i=0}^{10} \Pr(Heads|H_i) \times \Pr(H_i)$.

[29] Schoenfield (2019) recently argued in favour of a specific weighted version of Brier score, which does not only measure accuracy but also verisimilitude, i.e., how similar the probability distribution is to truth. Such a version should—contrary to the standard unweighted Brier score—be able to distinguish between the two above illustrative examples. We do not use this version here because the rule is relatively complicated and because a simple calculation of the distance between the expectation and objective probability of the outcomes suffices as a measure of verisimilitude for our present needs.

[30] Particularly, the simulations were conducted in $3 \times 11 \times 11 + 1 = 364$ variations (lying styles $\times$ inclinations to lie $\times$ coin biases + control run/ignoring the source).

biased towards landing on one side (e.g., the coins with 0.7, 0.8 bias towards one side). In other words, in case of the paradigmatic partial lies. These results can be interpreted by splitting lies into five types based on the source's degree of belief that the statement is false:

1. **Bullshit statements**
   The source is as certain that the statement is false as that it is true:
   $\Pr(\text{Statement is false}) = 0.5$
2. **Weak lies**
   The source is only weakly confident that the statement is false:
   $\Pr(\text{Statement is false}) \in (0.5, 0.6]$
3. **Medium lies**
   The source is somewhat confident that the statement is false:
   $\Pr(\text{Statement is false}) \in (0.6, 0.9)$
4. **Strong lies**
   The source is highly confident that the statement is false:
   $\Pr(\text{Statement is false}) \in [0.9, 1)$
5. **Strongest lies**
   The source is certain that the statement is false: $\Pr(\text{Statement is false}) = 1$

It should be noted that the ranges for weak, medium and strong lies we define here are arbitrary and are supposed to function as a helpful communicative tool. The spectrum between strongest lies and bullshit statements is continuous, and there is no precise boundary between the mentioned categories. However, the types of partial lying (i.e., 2.-4.) are helpful for our present needs because they allow us to interpret the results of our computer simulations. Weak lies are very close to bullshit statements, which means that they turn out to be false only slightly more often than true. Hence, the agent's level of trust τ quickly updates close to the level where the source is ignored (i.e., close to $\tau = 0.5$). In case of strong lies we observe a related phenomenon: the agent's trust quickly updates to a low level, so the source's statements are interpreted as almost the opposite of what is said. This results in a slight increase in overall epistemic inaccuracy. The medium partial lies, on the other hand, are the most damaging because the source's statements are relatively often true, yet, because of the medium level of distrust, they are interpreted as if they were mostly negations of what is said. This in the end results in an increase in epistemic inaccuracy (see Fig. 3 for a visualisation).[31]

In case the source is not constantly lying but is inclined to lie to some degree, we observe another interesting result, specifically, that the source needs to lie often enough to mislead the agent. This result is not surprising, however. Suppose the magician is flipping a coin that always lands tails. If he constantly lies, the player's level of trust will converge toward 0. In case of the weakest "lies" (i.e., "bullshit statements" when flipping a fair coin), the level of trust will converge toward 0.5 (except for clairvoyant lying) and to various levels between 0.5 and 1 for the other lies of in-between severity.

On the other hand, if the magician only lies sometimes—for instance in approximately 60% of the cases—then the agent's trust will converge toward 0.4 when the

[31] Only the results for coins with biases of 0.5 or more to land heads are represented because we obtain symmetric results with other biases as the situation is the same except that the magician makes the opposite statement ("heads" instead of "tails").

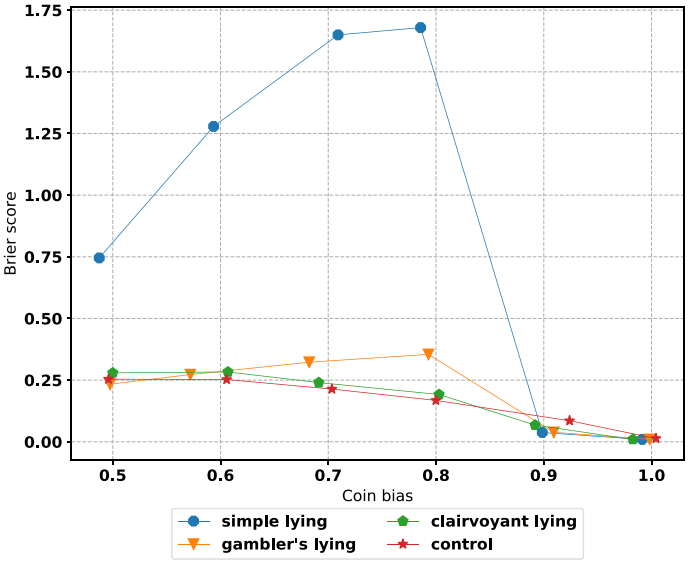

Fig. 3 Overall average Brier scores when being constantly lied to (over 1000 runs of 500 rounds); note that the lines in this and the following figures are visual aids only

magician is flipping a fully biased coin, towards 0.5 when he is flipping a fair coin and between 0.4 and 0.5 in the other cases. Because the player's level of trust will quickly stabilise towards 0.5 regardless of the severity of the lies, the magician will be largely ignored and the outcomes will approximate towards those of the control run, in which the agent only observes the actual outcomes of coin tosses.

This means that the severity of lying (as introduced with the framework mentioned above) loses its relevance when the source lies less often. The player is on average still somewhat less epistemically accurate in comparison to the control group but the Brier scores are correlated: in all cases where the source is less than 0.8 inclined to lie the overall inaccuracy increases when the coin is less biased. This includes all lying styles and the control runs. This is expected because it takes less time to identify a coin that is highly biased toward one side. The effect of lying at the same time decreases because the source is recognised as almost completely unreliable (for a similar notion of randomising unreliability see Bovens and Hartmann 2003, p. 57). This is clearly reflected in Fig. 4.

There is another interesting observation related to the epistemic damage that the simulations of lying according to various inclinations reveal. Recall that Krauss (2017) argued that we blame a person who is, say, 0.51 certain about something yet states it categorically, for the same reasons we blame liars (the expected epistemic damage suffered by a dupe). We believe these cases should not be classified as lying because the classification relies on the worse-off requirement, which is neither necessary nor sufficient to define lying (see Sect. 2 above and Benton 2018). However, if we simulate the magician as a truthful source—truthful in the sense that he always states the opposite of what he would state if he were lying in one of the three mentioned lying

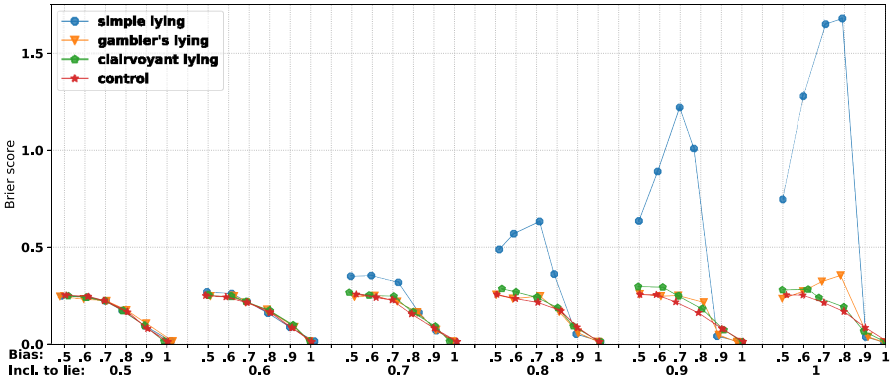

Fig. 4 The more constantly the source lies, the more pronounced the epistemic damage of being lied to is (overall average Brier scores over 1000 runs of 500 rounds)

styles—we obtain nearly the same results as when he is constantly lying. In case of full certainty that the statement is true (e.g., telling that a coin that can only land heads will land heads), the player improves her epistemic state. On the other hand, when the source is only somewhat certain about the outcome (e.g., a coin is 0.7 biased toward heads and he states that it will land heads), the updating proceeds in almost the same way as when he was lying. This is a consequence of how the likelihoods update in combination with the trust function. When a source is estimated to be partially lying about some outcome, his statements are interpreted in the same way as if the source was partially telling the truth. That is, if the level of trust, τ, is x in the former and $1-x$ in the latter case, then two opposite statements may be interpreted in the same way.

By means of an example: suppose the coin is 0.7 biased toward heads. If the source is (i) repeatedly partially lying that it will land tails, then τ will stabilise around 0.3 because the source tells the truth in approximately 30% of the cases. On the other hand, if the source is (ii) repeatedly partially telling the truth that it will land heads, τ will stabilise around 0.7. In (i), the dupe will interpret "Tails" as the source actually believing "Heads" with 0.7 probability, just like in case (ii) where the source says "Heads" but the addressee's $\tau = 0.7$. Similarly, the effect is more pronounced the more inclined the source is towards partial truth-telling. The outcomes for partial truth-telling are, therefore, analogous to partial lying; as expected this is also reflected in the results of the simulations. "Medium partial truth-telling" (in analogy with our classification) is, similarly, the most damaging partial truth-telling because it is hard to reliably use the trustworthiness estimate as additional information—just like in the case of medium lies. Hence, we can conclude that at least in the scope of our model, partial truths are just as epistemically damaging as partial lies, so one should only outright assert what one is highly confident to be true.

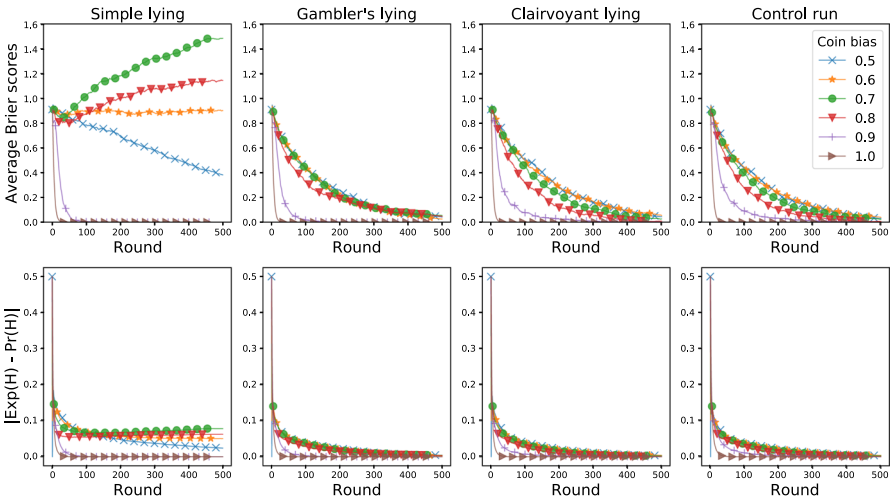

Fig. 5 Average Brier scores and the distance from the objective probability for the coin to land heads (over 1000 simulations) when the source is 90% inclined to lie and lies in the simple, gambler's and clairvoyant style and when the source is ignored (control run)

6 Should we ignore less than fully trustworthy sources?

The control runs (i.e., ignoring the source) typically led to the lowest overall Brier scores, or in other words, to the highest epistemic accuracy. This may indicate that as soon as one notices that the source is not completely trustworthy (in our case, by comparing the statements and observations of what side the coin landed on), a rational agent would do better to simply refrain from considering the source at all because she will very likely end up epistemically worse-off. However, epistemic accuracy is just one aspect an agent may be interested in; there are also other pragmatic considerations that need to be taken into account. For instance, it is pragmatically advantageous to be quicker than your peers in embracing true hypotheses (Douven 2013). As an example, consider two research teams that are investigating some phenomenon. Suppose both teams start researching the phenomenon at the same time. At some later point the findings of team A are still inconclusive in the sense that they are highly uncertain which hypothesis is true. Team B, on the other hand, is highly confident that they identified the true hypothesis, publishes their results in a prestigious journal and receives credit for the discovery. Although this is, admittedly, a gross simplification of scientific research, it shows that the speed of embracing true hypotheses is not irrelevant. This is similar to what happens when an agent updates her trust toward the source in the way we simulated in our model: it almost always leads to faster recognition of true hypotheses[32] (see, for instance, how the Brier scores decrease for the 0.9 biased coin in the plots of Fig. 5; similar plots for other variations and the raw data are available in Trpin 2020).

[32] By recognition (or embracing) of some hypothesis we mean that the agent assigns it high probability (in line with the probability norm of assertion; see, e.g., Lewis 1976, p. 303; Jackson 1979, p. 565, but also Douven 2006, pp. 457–459).

It may seem that it would be rational to ignore less than completely trustworthy sources. However, the additional information from an untrustworthy source may also benefit the agent in a more practical sense. Whether these advantages outweigh the epistemic risks of potentionally being misled remains an open question that cannot be categorically resolved or exhaustively discussed in this paper. Instead, it depends on the goals of the agent who finds herself in such a situation. That is, she needs to estimate the risks (overall greater epistemic inaccuracy) versus benefits (the speed of embracing the true hypothesis).

More specifically, we observe that in more than 70% of the cases not ignoring the source led to faster convergence toward true hypotheses in the sense that the agents were able to pass various thresholds of high probability[33] faster than if they ignored the information given by the source. However, this speed comes with a price: as noted, it may lead to a greater overall epistemic inaccuracy. This, in turn, means that the agent will sometimes assign very high probability to the wrong hypothesis (e.g., H_7 instead of H_8) and thus fail to assign high probability to the true hypothesis because the hypotheses are mutually exclusive and jointly exhaustive (there is only one coin used throughout the game). Even more so, we observe that if a lying style allows the agent to be much faster than the agent in the control group, this also increases the number of cases where the probability of the true hypothesis does not reach any of the thresholds.

For instance, the agent who is being lied to in the simple style on average passes the threshold of 0.99 for the true hypothesis over 100 steps[34] faster than the agent who ignores the source (control agent). However, the agent in the control group on average always passes the threshold of 0.95 probability for the true hypothesis in the first 500 rounds, while she fails to pass the threshold of 0.99 in 47.93% of the cases. The otherwise faster agent that is being lied to in the simple style, on the other hand, fails to pass even the 0.5 threshold of the probability for the true hypothesis in 19.01% cases and 58.08% of the cases for the threshold of 0.99 (see Table 4 for details). While the agents who ignore the source reach the highest threshold of 0.99 in the first 500 rounds most often, they are also the slowest in reaching it. This suggests that, as mentioned, there is a trade-off between speed and accuracy. Whether an agent will try to make use of the fact that she is being lied to should therefore also depend on agent's pragmatic goals (e.g., is it more important to embrace the true hypothesis faster or is it more important to be accurate).

7 Trusting by default

The results described so far show that partial lying, and medium lies particularly, may impose more epistemic damage on the dupe than categorical lies when the liar and

[33] For more details on the relevance and importance of various thresholds of high probability in similar cases see, e.g., Trpin and Pellert (2019, sec. 6), and Douven (2013, p. 433) in support of the threshold of 0.99 specifically.

[34] Each of the 500 rounds consists of two steps: the first step represents the update after the statement, the second after the observation. In case the source is ignored, the agent's probability distribution remains unchanged after the first step of the round.

Table 4 Percentage of simulations when a threshold probability for the true hypothesis was not passed and the average number of steps needed to pass various thresholds (in all variations for a given lying style)

	Control	Simple lying	Gambler's lying	Clairvoyant lying
Below 0.5	0.00%	19.01%	0.00%	0.00%
Below 0.75	0.00%	26.45%	0.00%	0.00%
Below 0.9	0.00%	34.71%	3.31%	0.00%
Below 0.95	0.00%	43.80%	6.61%	0.00%
Below 0.99	47.93%	58.68%	54.55%	50.41%
Steps to pass 0.5	91.47	103.08	90.93	86.06
Steps to pass 0.75	217.24	228.29	232.64	207.35
Steps to pass 0.9	401.49	337.24	405.30	390.72
Steps to pass 0.95	549.45	339.93	537.20	542.40
Steps to pass 0.99	525.81	299.80	377.78	491.80

The number of steps needed to pass higher thresholds is in some cases lower than that for the lower thresholds. This is because the numbers of steps are only calculated for the simulations where the threshold was passed

the dupe are in repeated interactions. It follows then that when the sequence of lies is longer, partial liars may—perhaps surprisingly—deserve more blame than categorical liars. This is because for the latter we can safely assume that they actually believe (almost) the opposite of what they state.

It should be noted, however, that we only considered the cases where the dupe immediately starts establishing how trustworthy the source is and takes this information into account. This seems not to be the case in everyday situations. There is sizeable empirical evidence from communication science and social psychology that people are truth biased, in the sense that we are not particularly good at recognising lies and that we tend to trust other people by default (see Levine 2014, for an overview of evidence and an original theory). As Levine argues, this behaviour might be seen as adaptively rational because most people are honest most of the time.

But what does this mean for our analysis, which was based on the assumption that the dupe is suspicious and establishes the trustworthiness of the source from the get-go? Would medium lies still turn out to be the most epistemically damaging if the simulated dupe was trustful by default, and would partial liars still deserve more blame than categorical liars? To address this potential worry we incorporated another variation of the standard scenario simulations, in which the dupe behaves more in line with the empirical evidence (e.g., the truth-default-theory by Levine 2014). The setup is the same: the player is trying to identify which coin (of the 11 described in Fig. 1 and Table 2) the magician is flipping. The lying styles and the inclinations to lie are also unchanged. The difference is in how the simulation progresses. The player is trustful by default. Hence, in the first (truth-default) phase at the beginning of the game she blindly trusts the magician with $\tau \geq 0.5$ (assuming bullshitting at worst). After this phase she becomes suspicious and ignores the magician ($\tau = 0.5$) (phase 2) until the real calculated level of trust τ falls below some distrust threshold. In this final phase

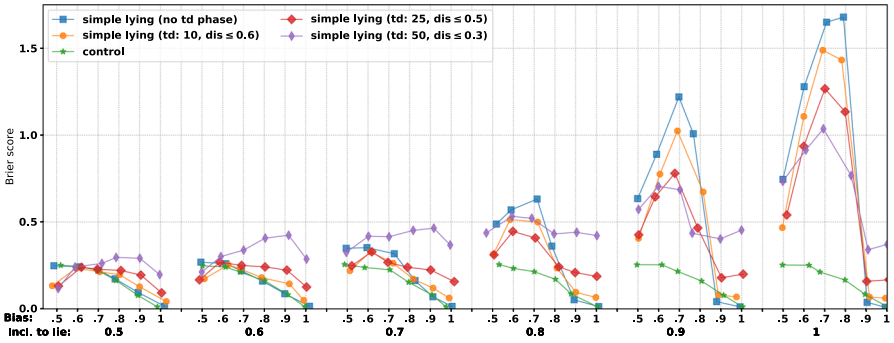

Fig. 6 Average overall Brier scores for simple lying according to different variations in comparison to the original scenarios and the control run (over 1000 runs of 500 rounds); td stands for the duration of the truth-default phase (number of rounds), dis for the distrust threshold

she then updates her beliefs just like in the original scenarios.[35] Note that throughout the game she also observes and updates her beliefs according to which side the coin lands on.

Incorporating this variation leads to a number of potential scenarios. We can vary the trust level that is invested into the magician in the first phase, the duration of the first phase, and the distrust threshold. As the number of simulations with new variations is exponentially increasing, we narrowed the variations down to the following:

- 4 different durations of the first phase (5, 10, 25 or 50 rounds)
- 4 different distrust threshold ($\tau < i, i \in \{0.3, 0.4, 0.5, 0.6\}$)
- 3 lying styles, and
- 6 different inclinations to lie (from 0.6 to 1 in 0.1 increments),

which led to $4 \times 4 \times 3 \times 1000$ (repetitions) = 360,000 additional simulations. The trust level in the first phase was randomised between 0.5 and 1 for each round.

To make the discussion more manageable, we focus only on 3 variations (the data and additional plots for other variations are available in Trpin 2020):

(a) Phase 1: 10 rounds, Phase 2: distrust if $\tau < 0.6$
(b) Phase 1: 25 rounds, Phase 2: distrust if $\tau < 0.5$
(c) Phase 1: 50 rounds, Phase 2: distrust if $\tau < 0.3$

The reason why we focus on these specific setups is that they represent three typical agents: (a) is suspicious and only trusts by default for a short time, then distrusts even if the source tells more truths than falsehoods; (c) is credulous and trusts by default for a long time, then only distrusts if the source tells considerably more falsehoods than truths. The agent (b) is in-between the two.

When we look at the results, we find a couple of interesting insights. The most important for our present needs is the fact that the partial lying is again more epistemically damaging than categorical lying, just as in the original scenarios. To be precise, categorical lying turns out to be slightly more damaging than strong partial

[35] We thank an anonymous reviewer for the suggestion of this variation inspired by social psychology and communication science.

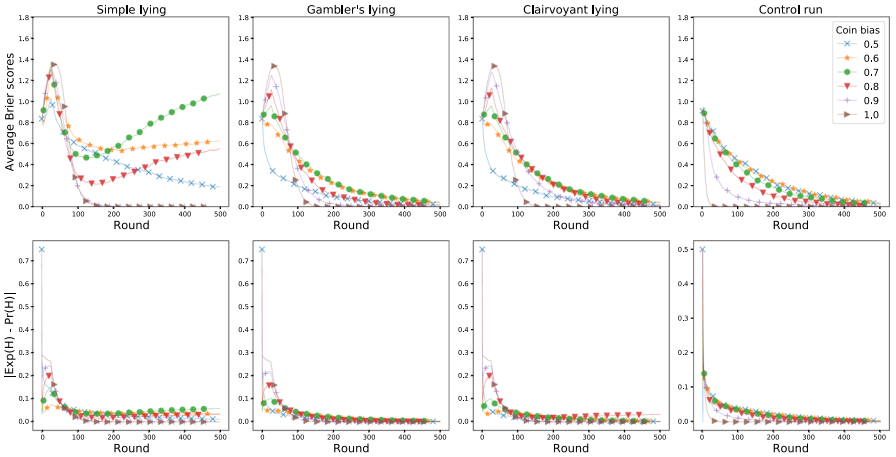

Fig. 7 Average Brier scores and the distance from the objective probability for the coin to land heads (over 1000 simulations) when the source is 90% inclined to lie and lies in the simple, gambler's and clairvoyant style and when the source is ignored (control run); the truth-default phase lasts for 25 rounds and the distrust threshold is 0.5

lying because it takes the player slightly longer to recover from the inaccurate state incurred in the initial trustful phase, in which the fact that the coin is fully biased becomes very unlikely (as a trusted source keeps stating to the contrary). However, medium lies again turn out to be the most damaging, as is the most clear in the case of simple lying (see Fig. 6; to improve legibility other lying styles are omitted in this plot but the full plot along with raw data is available in Trpin 2020).

There is also another observation that the new variation of the scenario uncovers, most clearly again in the case of simple lying: if the player is more credulous, then (in most cases) she will incur less epistemic damage compared to the other variations and the original scenario. This may strike us as surprising because the more credulous player exposes herself to the liar more than a suspicious player. However, further inspection reveals that the result is reasonable. For instance, suppose the coin has a bias of 0.8 to land heads and the magician lies in the simple lying style. The player may become increasingly confident that the coin is 0.5 biased during the initial trustful phase because the magician always tells her it will land tails, but she also observes a number of heads. Once she enters the final phase, her probability distribution updates and she considers various hypotheses: at first the 0.6 bias hypothesis seems the best, then 0.7, finally the 0.8 bias (true) provides a surprisingly good fit. However, because she brings her new trustworthiness assessment τ into play, she actually becomes more certain that the magician is lying than she should be, so the coin with the 0.9 bias (false) becomes the most viable hypothesis to which she slowly converges. Overall she therefore performs better than a suspicious player who misjudges the magician from the get-go because, despite the originally inaccurate phase, a phase in which she considers the true hypothesis (but ultimately fails to endorse it) emerges. This phase then brings the overall Brier score lower (see Fig. 7).

We will not go into more details related to these truth-default simulations like we did for the original scenarios because a thorough analysis (as well as potential further refinements and variations of the model) would exceed the scope of this paper. Our goal is, in the first place, to investigate the consequences of being partially lied to and to establish how lies of varying severity may affect the dupe. The truth-default variations thus primarily serves as a valuable robustness check for our main insight, which still holds: in the longer term partial lies, and medium lies specifically, impose more epistemic damage than categorical lies.

8 Conclusion

Our results provide an important insight in the debate related to the epistemic consequences of being (partially) lied to. Once longer sequences of interactions between a liar and a dupe, who adjusts her degree of trust towards the liar with respect to the statements that the liar is making, are considered, then the partial lies are generally more epistemically damaging than categorical lies because they are harder to detect. However, if the lying is not constant (i.e., if the liar is only more or less likely to lie in one of the senses described in this paper), the effect is less pronounced. If the lying is rare enough, the effect of being lied to almost completely disappears. The dupe is then treating the source to be neither reliable nor unreliable and simply ignores it.

Another important insight that the computer simulations reveal when compared to theoretical research on (partial) lying is that the consequences also differ in relation to the severity of partial lies. That is, if the partial lies are very strong in the sense that the liar is rightfully highly certain that a statement is false, then the lies are not as damaging. This follows since such lies are approximating the strongest possible lies (in the sense of asserting the opposite of what one is fully certain of) and the liar may be recognised as untrustworthy. Similarly, when the partial lies are weak, in the sense that the liar is only slightly more certain that the statement is false than that it is not, then the dupe is able to realise that the liar is only slightly more untrustworthy than trustworthy. Hence, the dupe is able to consider the liar to be largely irrelevant. However, when the liar operates with medium lies, that is the lies which are neither close to bullshit in Frankfurt (2005)'s sense nor close to strongest lies, then the dupe is able to see that the liar is reasonably untrustworthy. However, she cannot take this information into account as efficiently. Instead, she ends up in an epistemically worse situation when she tries to turn this into her advantage. It therefore seems that these situations—being the victim of medium lies—are the most dangerous and the dupe would be better off if she reflected her situation and immediately ignored the source once her degree of distrust is in this medium range. But the resolution is not so straightforward as taking the assessment of trustworthiness of the liar into account affords her to embrace the true hypothesis faster (if she does not end up embracing a false hypothesis). Therefore, there is no clear-cut winning strategy in situations where the source is considered to be at least to some degree untrustworthy; instead the dupe needs to consider her overall epistemic and pragmatic goals and decide for a strategy in line with these goals.

We chose to concentrate on the investigation of the dynamic relationship between partial lying and trust as little work has been done on this topic up to date, although the question is of high importance not just for epistemology but also for a wider understanding of human communication and cooperation in general (the recent catch phrases "fake news" and "post-fact world" are an illustrative case in point). While we acknowledge that simulation studies of elaborate psycho-social phenomena such as lying and trust strip away a lot of complexity associated with the real world interactions, computer simulations afford philosophical research a distinct advantage. Particularly, computational modelling and computer simulations allow for a systematical exploration of temporal dynamics of the cognitive processes of interest (here: trust dynamics and partial lies) and open a possibility for a more elaborate testing of intuitive predictions (Marsella et al. 2010) and more comprehensiveness in terms of evaluation of predictions (Fum et al. 2007), issues which otherwise suffer from small sample sizes or could hardly be investigated at all. Our research could thus be considered to be an exploratory step that provides grounds for further empirical research, which could focus on the specific case of medium lies and our ability to detect them. If they are indeed harder to detect than other types of lies, as our results suggest, then it seems reasonable that we should consider those who assert such lies often enough as more blameworthy than those who assert lies that are easy to recognise.

Although our goal is not to argue for external validation of our results but rather to propose a way of generating a viable empirical hypothesis, there is an aspect of our model that may strike one as crucially disanalogous to lying in real interactions. In our cases the agent is able to immediately see whether the magician's predictions are true or false. When the magician, for instance, makes a prediction "Heads", the agent afterwards observes what side the coin landed on and updates her trust towards the liar accordingly. If the prediction was correct, her trust increases, otherwise it decreases. The situation in realistic cases of repeated interactions, however, is often different: the truth or falsehood of many statements may only be verified much later (or perhaps never). An argument could therefore be made that our results are only relevant for very specific cases where an agent can immediately and repeatedly verify the source's statements.[36]

Nevertheless, we anticipate that the outcome in such situations would typically not be importantly different to the results we presented in this paper. Because the agent uses Laplace succession rule as a way of probabilistically updating trust towards the source this means that the agent only gradually moves to a level of trust where the liar's statements have larger impact (e.g., after 10 verifications of statements). The agent would therefore ignore the source for a longer initial period in case of unverified (or unverifiable) statements. A bigger problem would arise in cases where the source would initially establish agent's high level of trust (by repeatedly making verifiably true statements or because the agent is trustful by default). The source could at that point start asserting false statements, which the agent could not verify (or could only verify much later). The dupe in our model would in this case end up in a more inaccurate epistemic position as her level of trust would remain very high until the statements

[36] We would like to thank an anonymous reviewer for bringing up this potential objection.

can be recognised as false. Whether similar cases are common in real interactions is another question that may be addressed by empirical research.

There are still two responses we can state at this point. First, it seems that this is not only a problem for our model but also for real-world interactions—a smart devious liar will make sure to establish trust with the dupe before lying as the consequences of lying will thus be more devastating for the dupe. Second, the way we model the dynamics of trust is admittedly simplified. However, we could account for cases like this by introducing a more complex way of updating trust. We leave these questions for future research.

In addition to that, we believe that the results could find place in other debates that are not related to lying as such but rather to the general concepts of (un)reliability and biased sources, for instance in scientific reasoning (see, e.g., Osimani and Landes 2020). We are therefore looking forward to new research that may in some way benefit from our paper: be it by further investigating the concept of partial lies, the concept of graded trust, general problems of (un)reliability under uncertainty or in some other way which we do not yet foresee.

Acknowledgements Open Access funding provided by Projekt DEAL. This paper benefited from the questions and comments of the audiences at the Communication and Cooperation Workshop (University of Milan), the MEi:CogSci Conference (University of Bratislava), and the Computational Modeling in Philosophy Conference (Munich Center for Mathematical Philosophy) where we presented the preliminary results in June 2018. We would also like to thank Anton Donchev and three anonymous reviewers of this journal for their helpful comments on earlier versions of the paper. Parts of the research (B.T.) were funded by Ernst Mach Grant, Ernst Mach Worldwide (ICM-2018-10093; University of Salzburg), and by Alexander von Humboldt Foundation (LMU Munich). The project initially started in the scope of the MEi:CogSci programme at University of Ljubljana.

Author contributions The first author wrote the paper, the code used for the computer simulations and designed the overall study. The second and the third author are listed in alphabetical order. They equally researched the literature, contributed to the overall directions of the study and improvements of the paper.

References

Benton, M. A. (2018). Lying, accuracy and credence. *Analysis*, *78*(2), 195–198.

Bovens, L., & Hartmann, S. (2003). *Bayesian epistemology*. Oxford: Clarendon Press.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78*(1), 1–3.

Cosmides, L., & Tooby, J. (1992). Cognitive adaptations for social change. In J. Barkow, L. Cosmides, & J. Tooby (Eds.), *The adapted mind: Evolutionary psychology and the generation of culture* (pp. 163–228). New York: Oxford University Press.

Douven, I. (2006). Assertion, knowledge, and rational credibility. *The Philosophical Review*, *115*(4), 449–485. https://doi.org/10.1215/00318108-2006-010.

Douven, I. (2013). Inference to the best explanation, Dutch books, and inaccuracy minimisation. *Philosophical Quarterly*, *63*(252), 428–444.
Frankfurt, H. G. (2005). *On bullshit*. Princeton, NJ: Princeton Univerity Press.
Fum, D., Missier, F. D., & Stocco, A. (2007). The cognitive modeling of human behavior: Why a model is (sometimes) better than 10,000 words. *Cognitive Systems Research*, *8*(3), 135–142. https://doi.org/10.1016/j.cogsys.2007.07.001.
Hartmann, S., & Sprenger, J. (2010). Bayesian epistemology. In S. Bernecker & D. Pritchard (Eds.), *Routledge companion to epistemology* (pp. 609–620). London: Routledge.
Jackson, F. (1979). On assertion and indicative conditionals. *The Philosophical Review*, *88*(4), 565–589.
Krauss, S. F. (2017). Lying, risk and accuracy. *Analysis*, *77*(4), 726–734.
Leitgeb, H., & Pettigrew, R. (2010a). An objective justification of Bayesianism I: Measuring inaccuracy. *Philosophy of Science*, *77*(2), 201–235.
Leitgeb, H., & Pettigrew, R. (2010b). An objective justification of Bayesianism II: The consequences of minimizing inaccuracy. *Philosophy of Science*, *77*(2), 236–272.
Levine, T. R. (2014). Truth-default theory (TDT): A theory of human deception and deception detection. *Journal of Language and Social Psychology*, *33*(4), 378–392.
Levinstein, B. A. (2012). Leitgeb and Pettigrew on accuracy and updating. *Philosophy of Science*, *79*(3), 413–424.
Lewis, D. (1976). Probabilities of conditionals and conditional probabilities. *The Philosophical Review*, *97*(4), 497–520.
Lorini, E., & Demolombe, R. (2008). From binary trust to graded trust in information sources: A logical perspective. In R. Falcone, S. K. Barber, J. Sabater-Mir, & M. P. Singh (Eds.), *Trust in Agent Societies* (pp. 205–225). Berlin: Springer.
Mahon, J. E. (2016). The definition of lying and deception. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (winter 2016 ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2016/entries/lying-definition/.
Marsella, S., Gratch, J., & Petta, P. (2010). Computational models of emotion. In K. R. Scherer, T. Bänziger, & E. B. Roesch (Eds.), *Blueprint for affective computing: A sourcebook* (pp. 21–46). Oxford: Oxford University Press.
Marsili, N. (2014). Lying as a scalar phenomenon. In S. Cantarini, W. Abraham, & E. Leiss (Eds.), *Certainty-uncertainty—And the attitudinal space in between*. Amsterdam: John Benjamins Publishing.
Marsili, N. (2018). Lying and certainty. In J. Meibauer (Ed.), *The Oxford handbook of lying*. Oxford: Oxford University Press.
Montaigne, M. (1910). *Essays of montaigne* (Vol. 1). New York: Edwin C. Hill.
Osimani, B., & Landes, J. (2020). Varieties of error and varieties of evidence in scientific inference. *The British Journal for the Philosophy of Science* (Forthcoming).
Robert, C. (2007). *The Bayesian choice: From decision-theoretic foundations to computational implementation*. Berlin: Springer.
Rothschild, D. (2020). What it takes to believe. *Philosophical Studies*, *177*(5), 1345–1362. https://doi.org/10.1007/s11098-019-01256-6.
Schoenfield, M. (2019). Accuracy and verisimilitude: The good, the bad and the ugly. *The British Journal for the Philosophy of Science* (Forthcoming).
Sorensen, R. (2007). Bald-faced lies! Lying without intent to deceive. *Pacific Philosophical Quarterly*, *88*(2), 251–264.
Trpin, B. (2020). Raw data and additional plots for the paper Lying, more or less: A computer simulation study of graded lies and trust dynamics. *Mendeley Data*. https://doi.org/10.17632/8bpzprrp4y.1.
Trpin, B., & Pellert, M. (2019). Inference to the best explanation in uncertain evidential situations. *The British Journal for the Philosophy of Science*, *70*(4), 977–1001. https://doi.org/10.1093/bjps/axy027.
Vallinder, A., & Olsson, E. J. (2014). Trust and the value of overconfidence: A Bayesian perspective on social network communication. *Synthese*, *191*(9), 1991–2007.