Philosophia
https://doi.org/10.1007/s11406-020-00191-9

Kant Can't Get No . . . Contradiction

Neven Sesardić[1]

Received: 15 July 2019 / Revised: 15 July 2019 / Accepted: 5 February 2020
Published online: 23 March 2020

Abstract
According to Kant, the universalization of the maxim of false promising leads to a contradiction, namely, to everyone adopting the maxim of false promising which would in effect make promising impossible. I first propose a reconstruction of Kant's reasoning in four steps and then show that each of these steps is highly problematic. In the second part I argue that attempts by several prominent contemporary philosophers to defend Kant fail because they encounter similar difficulties.

Keywords Universalization · Contradiction · False promises · Kantian ethics

Kant famously claimed that universalizing the maxim of false promising leads to a contradiction: if everyone accepted the maxim behind false promising, the trust in promises would completely disappear and the institution of promising would break down. Whichever way the contradiction is supposed to be derived from the universalization (i.e. whether the outcome is seen as a "contradiction in conception" or "contradiction in the will"), the main source of the contradiction is the disappearance (or even impossibility) of promising that is presumed to be the consequence of the universalization. Roughly, how could one promise (falsely) in a world in which the institution of promising no longer existed?

I will argue, however, that Kant failed to show that the universalization of the maxim of false promising leads to the breakdown of promising. In the first part I will try to show that when Kant's reasoning is laid out in full and carefully scrutinized, it will become clear that it contains four inferential steps, each of which lacks adequate justification. In the second part I will give examples of prominent contemporary philosophers who followed Kant and uncritically accepted at least some steps of his highly problematic argument.

✉ Neven Sesardić

[1] Zagreb, Croatia

1 Kant on False Promises

Kant's argument about false promising has been described as his best example (Parfit 2011: 279; Parfit 2004: 320; Korsgaard 1996: 14), the most helpful example (Maclntyre 1998: 124), the easiest example (Korsgaard 1996: 63), one of his most persuasive examples (Blackburn 2001: 121; Frankena 1973: 31), and an example "where there is a single clear and decisive answer" (Wood 2008: 65).

Suppose I need money and try to get a loan by falsely promising to repay it later. Could I will that everyone behaved in the same way in similar situations? Kant says no:

> …how would it be if my maxim became a universal law? I then see at once that it could never hold as a universal law of nature and be consistent with itself, but must necessarily contradict itself. For, the universality of a law that everyone, when he believes himself to be in need, could promise whatever he pleases with the intention of not keeping it would make the promise and the end one might have in it itself impossible, since no one would believe what was promised him but would laugh at all such expressions as vain pretenses. (G 4:422; Kant 1997)

And similarly:

> …could I indeed say to myself that every one may make a false promise when he finds himself in a difficulty he can get out of in no other way? Then I soon become aware that I could indeed will the lie, but by no means a universal law to lie; for in accordance with such a law there would properly be no promises at all, since it would be futile to avow my will with regard to my future actions to others who would not believe this avowal or, if they rashly did so, would pay me back in like coin; and thus my maxim, as soon as it were made a universal law, would have to destroy itself. (G 4:403)

If the maxim behind a lying promise were universalized, it is claimed, promising would become impossible because no one would take any promise seriously.

To get a better picture of the Kantian reasoning, let us present it as a sequence of six steps: two preparatory steps and four stages of the argument proper (see Table 1). The parts in bold are my additions that go beyond what can be found in Kant's text. These additions are introduced to identify and fill the gaps in the inferential chain while at the same time hopefully being true to the spirit of Kant's argument.

The steps in Table 1 are successive stages of Kant's reasoning, according to the most charitable interpretation I could muster. The first two steps, *M* and *U*, set the stage: they describe the maxim behind the lying promise and its universalization. The argument then proceeds further in four steps (*~R*, *B~R*, *~T,* and *C*).

Let us now take a closer look at each of the six steps.

(M) The maxim behind false promises

This is how Kant describes the maxim behind a lying promise: "When I believe myself to be in need of money I shall borrow money and promise to repay it, even though I know that this will never happen" (4:422).

Table 1 Reconstruction of Kant's argument

Step		Explanation/justification
M	Whenever I am in need I will make a false promise **when it is in my interest to do so**.	The maxim underlying a false promise.
U	It is a universal law of nature that everyone accepts maxim *M* and acts on it when the conditions are satisfied.	Universalization of maxim *M*.
~R	**There is no good reason to trust promises.**	Via *U*
B~R	**Everyone believes that ~R.**	Via *~R*
~T	No one will trust promises any longer[a].	Via *B~R*
C	The practice of promising crumbles[b]. It becomes impossible to promise.	Via *~T*

[a] In Kant's own words: "…no one would believe what was promised him but would laugh at all such expressions as vain pretenses."

[b] Kant: "…it would make the promise . . . itself impossible," and "there would properly be no promises at all."

The problem with Kant's version of the maxim is that, taken literally, it too rigidly links being in need of money with false promising. But it is simply not true (nor did Kant take it to be true) that people, *just because they need money*, try to get a loan from others by falsely promising to repay it. They use that strategy *only if* they think this is in their interest, *all things considered* (e.g. they don't give false promises to people who are unlikely to give them loans, nor to dangerous criminals who are likely to take revenge on them later, etc.). My friendly amendment[1] "when it is in my interest to do so" is in line with Kant's own observation that false promisers take the long view and do not look only at immediate consequences: "I see very well that it is not enough to get out of a present difficulty by means of this subterfuge but that *I must reflect carefully whether this lie may later give rise to much greater inconvenience for me than that from which I now extricate myself.*" (4:402—emphasis added)

(U) The universalization of (M)

This step (universalization) is uncontroversial. It just assumes that it is a universal law that everyone adopts *M* and acts on it under the specified circumstances.

(~R) There is no good reason to trust promises

When Kant says that in a *U*-world "no one would believe what was promised him but would laugh at all such expressions as vain pretenses," it is not clear how exactly he arrives at that conclusion. The most plausible interpretation, I think, is that he relies on something like *~R*. But why would *U* (the universalization of maxim *M*) result in *~R* (the statement that there is no good reason to trust promises)?

One might wonder whether when Kant speaks of *U* as "the universal law to lie" he perhaps thinks that in a *U*-world *no one would ever intend to keep any promise*. If the italicized sentence were true, then *U* would indeed imply *~R* (that there would be no

[1] Other commentators also realize that Kant's maxim has to be modified in a similar way: e.g. Maclntyre 1998: 124; Singer 1998; Wood 1972: 618; Rawls 2000: 170; Guyer 2014: 221.

good reason to trust promises). But it does *not* follow from *U* that no one would ever intend to keep any promise, for (as explained above) *M* should *not* be interpreted as saying "I will *always* break my promises, *no matter what*" but only "I will break my promise *whenever I think this would be in my interest*." This maxim obviously leaves ample room for one to (intend to) keep some (many?) of one's promises, namely in those situations in which one would benefit from this. Even in Kant's example of promising to return the borrowed money it will often be in the borrower's interest to keep the promise (say, if he can hope to lead the lender on by repaying him a small loan with high interest, creating trust, then later borrowing a much larger sum from him and finally absconding with the money). Obviously, then, not every individual adopting maxim *M* would intend to break all his promises in a robot-like fashion and without a single exception.[2] Some promises could still be honored (and trusted) even if everyone accepted and acted on maxim *M*. Therefore, the universalization of *M* by no means entails that no one would ever intend to keep any promise.

But what other justification could be given for inferring ~*R* from *U*? One could argue: if people kept their promises only when this was in their interest, wouldn't this destroy any sufficient reason for trusting promises simply because it would be very hard to tell the difference between sincere and deceitful promises? Actually no, because, first, "very hard" does not mean "impossible." And second, even if it were *impossible* to distinguish between sincere and false promises, it might still be reasonable to trust *all* promises. Suppose that only 20% of promises were deceitful and that people could not distinguish them from honest promises. Then confronted with a promise (and with no other information) one would be entitled to trust it with the relatively high probability of .8. So, ~*R* would be false.

Furthermore, it is sometimes quite easy to distinguish between the two kinds of situations, i.e. when trust would be justified and when not. For example, if X promises to repay a loan to Y, Y will have a good reason to trust X if X has a stable and highly attractive job in which the reputation for honesty is very important. However, Y will not have the same level of trust in Z, a stranger whom he just met in a bar and who made the same promise. So in a world in which *U* holds, and everyone knows it, one would expect that people would try harder to get more information about the situation of a promiser before giving him a loan, because they would be aware of a higher probability of insincere promises. But it wouldn't be generally true that there would be no good reason to trust promises. It would depend on the situation.

Nevertheless a worry may linger: how could I trust anyone in a *U*-world if I know that everyone in that world who does intend to keep his promise to me may always change his mind if he decides that keeping the promise is not actually in his interest? Yes, he may change his mind, but this mere possibility should not make me lose trust in his promise as long as I have reasons to believe that this is *unlikely* to happen.

All in all, it is not just that *U* does not justify ~*R*. It is unclear that there is a plausible argument that would show that in a *U*-world we should necessarily expect ~*R* to be true.

[2] This is why Fred Feldman is wrong when he says that if the maxim of false promising were universalized then "no one could borrow money on a false promise - for *if promises were always violated*, who would be silly enough to loan any money?" (Feldman 1998: 193—emphasis added). If the maxim of false promising were universalized, it is *not* true that promises would be *always* violated.

(B~R) Everyone believes that ~R

Although Kant himself neither introduced nor explicitly discussed anything that would match *B~R*, it is nevertheless inserted into the table representing the reconstruction of his argument. Why? Simply because *B~R* is indispensable for moving the argument forward from *U* and *~R* toward establishing *C*.

The only road from *~R* to *~T* goes via *B~R*. From the mere truth of *~R* (that there is no good reason to trust promises) it does not follow that no one would trust promises (*~T*). To draw that conclusion an additional premise is needed, namely that *all people are also aware* of the truth of *~R*. (Simply, a universal distrust of promises will not be brought about merely by promises being unreliable, but only if, in addition, all people also *believe* that promises are unreliable.)

Now what might be a justification for *B~R*? It certainly cannot be *~R* alone. The statement "*~R* is *universally* believed to be true" contains a very strong and highly implausible claim that cries out for justification before it is admitted as a premise. And if the statement were weakened by the word "universally" being replaced with "mostly" or "typically" it would gain on plausibility but it could no longer bear the brunt of supporting *~T.* For as long as there are at least some people ignorant of the truth of *~R*, they could be duped by a cunning false promise and so it wouldn't be true that "*no one* trusts promises any longer".

To be fair to Kant, there is no doubt that the truth of *~R* would decrease the level of trust in promises considerably. For if all promises were untrustworthy, many people would soon take notice and become more reluctant to trust promises. But even in a world in which *~R* were true and this were a consequence of a law of nature (*U*), we should not assume that this would lead to *B~R,* and that consequently the trust in promises would disappear completely. For how many true laws of nature (or their consequences) are there that are believed (or known) to be true by *everyone*? The correct answer is probably "none". Hence why would *~R* be an exception?

Despite it being a universal law that it is impossible to communicate with the dead, some people not only believe that they can receive messages from their deceased loved ones but are also prepared to pay a lot of money to those who promise to help them make contact with the great beyond. With such a level of human credulity it is hard to imagine that everyone would believe *~R* even if *~R* were a universal law of nature (which it is not in Kant's argument).

Consider children living in a *~R*-world. As they grow up they will obviously need some time to learn the ugly truth about the spread of human dishonesty. So, as long as there are children who are not yet weaned off their naïve and immature trust in promises (and many such children will always be around), there will be an ever fresh supply of "suckers" who will make it possible for the practice of giving and accepting promises to go on (despite the truth of *~R*). Therefore, contra Kant, the trust in promises would not completely disappear.

And it is not only about children. As in the actual world, many of those issuing lying promises in a *~R*-world would try to preserve some credibility by keeping their promises occasionally, or concocting retroactive excuses (e.g. "I did sincerely intend to pay back the debt but unanticipated events unfortunately precluded this"), or by giving some evidence that their debt has been settled (e.g. by showing documents that look like bank confirmations of money transfers), or by making up other stories to "prove" their good faith, etc. Now it is highly unlikely that *all* these excuses of *all* these promisers would be resolutely rejected by *everyone* to whom this kind of excuse would *ever* be presented.

Human gullibility is again the key. Take the famous case in which some people were ready to sell all their possessions, leave their families and join a sect because they swallowed the story of a Chicago housewife that the end of the world was coming and that they could be saved only by following her advice and meeting at a certain place to board a flying saucer that would take them to safety to the faraway planet Clarion (Festinger et al. 1956). It would be amazing if absolutely no one in a ~*R*-world would fall for some much less far-fetched lies, like the one about promising to return a small amount of the borrowed money.

If in our world there are still people who believe that the Earth is flat, why wouldn't at least some in a ~*R*-world also believe (wrongly) that promises are sometimes trustworthy? And as long as there are people around who would take the bait, one could—contrary to what Kant says—issue a lying promise and without any contradiction will that one's maxim be universalized. For even in a world in which no promises are trustworthy, one could still reasonably hope to find people whom one could cheat with a false promise.

(~T) No one will trust promises any longer

Would granting *B*~*R* be sufficient to establish ~*T*? In some sense, not quite yet. Imagine a world in which both ~*R* and *B*~*R* are true: there is no reason to trust promises, and also everyone happens to know that. Now is it possible that ~*R* and *B*~*R* are both true at the time but that ~*T* ("No one will trust promises any longer") is still false? In other words, would it be possible, in the "~*R* & *B*~*R*" world, to plan to issue a promise and hope that someone would trust it?

Well, although we assume that everyone there *at the time* believes (correctly) that there is no reason to trust promises, people's beliefs are open to change and so there is no principled reason why even in that world someone could not manage to invent a new trick and convince some people (falsely) that, exceptionally and remarkably, this particular promise is sincere.

Notice the difference in status between propositions ~*R* and *B*~*R* in Kant's argument. In his thought experiment ~*R* is supposed to be a consequence of a *natural law*, whereas *B*~*R* is at best a statement that is regarded as being *contingently* true, i.e. something that is an expected—but *not* necessary—consequence of ~*R*. So if *B*~*R* happens to be true at t_1 it doesn't have to stay that way at t_2.

People occasionally withdraw their claims to knowledge (not always with good reason), so this can also happen in a world in which *B*~*R* is true at a given time. This explains how, contra Kant, one's intention to cheat someone with a false promise in a world in which both ~*R* and *B*~*R* happen to be true at a moment can in fact be perfectly coherent. One may come up with a specially devious plan that one hopes will trick some people and thereby turn the truth value of B~R from true (at t_1) to false (at t_2). And one's plan could work, without generating a contradiction.

(C) The practice of promising crumbles

Finally, let us consider the last step in Kant's reconstructed argument, inferring *C* (that the practice of promising will crumble) from *~T* (that a trust in promises permanently disappears).

At first it might seem that Kant's argument is here unassailable. For if the trust in promising were completely extinguished, wouldn't everyone then just "laugh at all such expressions as vain pretenses"? Isn't it clear that under these circumstances any attempt to borrow money by promising to return it later would be completely pointless and would soon disappear from the human behavioral repertoire, as no one would trust anyone else's promise?

Maybe not. Person A could have excellent reasons to lend money to B even when A could not rely on B to keep his promise to repay the loan. One way this could work is if A were in a position to issue a credible threat to B in case B didn't return the money and thereby force him to change his mind and become more cooperative. Or, alternatively, A could do what many people in our world already do, i.e. make sure he has solid evidence of B's making the promise and then hope to recover the debt by suing B and enforcing payment via a court order. So even if it were very likely that a promiser had no intention to keep his promise and if the promisee were completely aware of that fact, the act of promising-cum-borrowing could still take place as long as the promisee believed he had a good chance of enforcing the payback of the loan (compounded with interest) through some kind of legal system.[3] In a word, even the universal untrustworthiness of promising plus the universal awareness of that untrustworthiness would not necessarily destroy the practice of promising.

Kant thought that not trusting a potential borrower would be a sufficient reason not to lend him the money. But this seems wrong. The fact that a would-be borrower makes a promise with the sole intention to break it and fleece the lender does not mean that his scheme will work. The lender may have no trust whatsoever in the borrower and yet reasonably decide to lend him the money in the hope that he will out-scheme the schemer.

Such an activity already exists in our world. There are people who make it their business to lend money precisely to those whom they suspect of being unwilling to repay the debt. The lenders may nevertheless be confident that they will manage to recover the money (plus a profit from the late payment interest) by putting pressure on the borrowers through intimidation, beatings, threat of legal action, etc. So a complete lack of trust in a promiser is not necessarily an obstacle to handing him over money on the basis of a promise.

Now if we imagine that the practice of false promising in our world starts spreading until eventually *~R, B~R,* and *~T* all become true, there is no reason why in that world some people would not continue to use others' recognizably untrustworthy promises as

[3] But wouldn't that legal system, too, rely on trust in promises (e.g. to uphold the law)? Not necessarily. It is not clear why it would be impossible to set up a system according to which anyone caught reneging on a promise would be punished and forced to compensate the promisee. Now why would such a system be introduced? Well, cooperation is socially important, so if people could not be expected to keep their promises out of sense of duty, an obvious alternative would be to make them do it out of fear of punishment (with the additional arrangement that punishers are rewarded and non-punishers punished, which would add stability to the system).

an opportunity to earn money through risky lending and consequent attempts to enforce loan repayment (with high interest). Therefore it seems wrong to believe that promises would necessarily disappear or become impossible in that world.

But would these still be promises, really? There are two answers to this question: "Yes" and "It doesn't matter."

First, according to Merriam-Webster the noun *promise* has two primary meanings: (a) "a declaration that one will do or refrain from doing something specified", and (b) "a legally binding declaration that gives the person to whom it is made a right to expect or to claim the performance or forbearance of a specified act." So the acts in a ~*T*-world that I called "promises" do qualify as promises by the dictionary definition of the term. Hence those philosophers who try to defend Kant by insisting that "a promise that no one would believe would not count as a promise" (Allison 2011: 186) do not appear to have ordinary language on their side. Going deeper into the meaning and status of promising, it is far from clear that the requirement "X is a promise *only if* it would be believed by someone" is built as a constitutive rule into main theoretical approaches to promising (see the overview in Vitek 1993).

But second, the answer doesn't really matter because even if it were shown that the acts in question are not strictly promises this would not really help Kant's argument. For, his point was not a verbal but a substantive one. He was not interested in how a lying promise should be described in a ~*T* world (i.e. whether it should be called a real promise or quasi-promise). Rather, he wanted to prove that by issuing what we call a lying promise in our world, a "promiser" could not achieve his goal of extracting money from a "promisee" in the world in which there was a universal lack of trust in promises. And, as we have just shown, he did not prove that.

The key thing to keep in mind here is that if X lends money to Y on the basis of Y's promise to return it, this does not entail that X trusts Y. Here is another kind of situation that illustrates that promising is not so rigidly connected to trust. Suppose that in a ~*T* world (in which people don't trust promises) a psychologist conducts a study with the aim of finding out how low the frequency of fulfilling promises[4] is under certain circumstances. He approaches a number of randomly chosen subjects and offers them an arrangement whereby a loan to them gets automatically approved and the money immediately transferred once they promise to return the money within a certain period. The study is published reporting the very low percentage of cases in which the loan was returned. Replications follow.

This example shows, again, that, contra Kant, giving a loan on the basis of a promise to return it *could continue to exist even in a world in which no one would trust promises*. The reason is that giving a loan on the basis of a promise does not have to be based on trust. There can be other motives (in this case, scientific curiosity).

This completes the first part of this article that points to serious weaknesses in Kant's argument at every one of its four stages. Let us now look at how his argument has been treated by some of the leading contemporary philosophers.

[4] A promise can be fulfilled either because (a) it was an honest promise, or (b) because it was fulfilled by mistake, confusion, insanity etc. Ad (a): notice that ~*T* by itself does *not* strictly imply that there are never any honest promises, only that *if some exist*, they are rare and no one could tell them apart from much more frequent lying promises (and for this reason no one would trust them). Ad (b): obviously, the situation in which people would "honor" their promises by glitch would be perfectly possible in a ~*T* world.

2 Contemporary Philosophers on Kant on False Promising

2.1 Alasdair MacIntyre

> If all men acted upon this precept [“I may always break a promise when it is in my interest to do so”], and broke their promises whenever it suited them, clearly the practices of making and of relying upon promises would break down, for nobody would be able to trust the promises of others, and consequently, utterances of the form “I promise to . . .” would cease to have point. (Maclntyre 1998: 124)

Macintyre starts with “If all men acted upon this precept…” (our *U*), and says that this would “clearly” lead to the breakdown of the practice of promising (our *C*) via the proposition that “nobody would be able to trust the promises of others” (our ~*T*).

MacIntyre’s direct derivation of *C* from *U* is too quick. It should include the intermediate steps of establishing the truth of ~*R* (presumably deriving it from *U*), and then deriving *B*~*R*, ~*T* and *C*. However, the prospect of accomplishing all these tasks seems bleak.

Take someone who accepts MacIntyre’s precept “I may always break a promise *when it is in my interest to do so*.” Can I trust that person’s promise? It depends. The precept obviously allows for the possibility that he sometimes gives honest promises, namely when giving an honest promise *is in his interest*. Then I should just make sure that he makes a promise to me in a situation that seems to be of that kind and I could have some level of trust in his promise. Therefore, there is no reason to accept MacIntyre’s conclusion that, with the universalization of the maxim of false promising, the practice of promising would “clearly” break down. There is nothing clear about that.

2.2 Onora O’Neill

> A maxim of deceit can readily be seen as one that we cannot even conceive as universally adopted. The project of deceit requires a world with sufficient trust for deceivers to get others to believe them; the results of universal deception would be a world in which such trust was lacking, and the deceiver's project was impossible… There is a fair amount of agreement on this type of account of contradictions in conception, but little on whether it provides a plausible criterion for the range of duties conventionally classified as perfect duties. (O'Neill 1989: 132)

Although O’Neill is right that “there is a fair amount of agreement” with the claim that the universalization of the lying promise maxim leads to a contradiction, its many problems should be familiar by now. First, the universalization of a “maxim of deceit” (*U*) does *not* lead to universal deception. Remember, the maxim is “I will make lying promises *when it is in my interest to do so*”, so even if everyone adopted that maxim, this would still be perfectly compatible with people often making promises they intend to keep. Second, even universal deception would not by itself produce a world with no trust in promising (*~T*). (*B~R* would be needed for this, and *B~R* is implausible.) Third, the lack of trust (*~T*) would not necessarily lead to the deceiver's project being impossible or to the breakdown of promising (*C*).

O'Neill is also wrong when she claims that if a deceitful promise is accepted, then "the person to whom it was given must be ignorant of what the promisor's intention (maxim) really is." She explains elsewhere:

> If one knew that the promisor did not intend to do what he or she was promising, one would, after all, not accept or rely on the promise. It would be as though there had been no promise made. (O'Neill 1994: 44)

No, a promisee *could* accept a promise, despite knowing that the promiser does not intend to keep it, because even deceitful promises have consequences. They do give rise to obligations even if the promisee is aware of the deceit. It would *not* be "as though there had been no promise made."

2.3 Christine Korsgaard

> A man who wills to use the institution of promising in pursuit of his end must will that the institution should work. And it does not work unless promises are generally made in good faith. (Korsgaard 1996: 64)

But why couldn't the institution of promising work even if promises were *not* generally (i.e. usually, as a rule) made in good faith? Imagine a world in which it was universally known that less than 1% of promises were given in good faith. Wouldn't at least some people in that world try to identify—and perhaps make a business deal with—those few trustworthy promisers, whereby the institution of promising would carry on, despite the massive spread of dishonesty? After all, why wouldn't this be possible if nowadays there are still people who place confidence even in those promises that are very widely known *never* to have been given in good faith, e.g. the promises of receiving millions of dollars from an African prince if you help him withdraw money from a bank account to which he has allegedly lost access? This notorious Nigerian email scam has *not* "died out under stress of too many violations" (Korsgaard's phrase, see below), despite the fact that only a small minority of people still fall for it. (According to a report released by Ultrascan AGI, losses from Nigerian scams were over $12 billion in 2013.)

This case is actually very instructive for us because these scammers face the same kind of problem that false promisers would have to solve in a world in which very few people would fall for false promises. And the Nigerian email scammers in fact found a solution, which is not immediately recognized as such. Many people laugh at the crudity of the scam but this is only because they don't understand how it works. Basically, what looks like a bug is really a feature. Microsoft researcher Cormac Herley explains:

> Far-fetched tales of West African riches strike most as comical. Our analysis suggests that is an advantage to the attacker, not a disadvantage. Since his attack

> has a low density of victims the Nigerian scammer has an over-riding need to reduce false positives. By sending an email that *repels all but the most gullible* the scammer gets the most promising marks to self-select, and tilts the true to false positive ratio in his favor. (Herley 2012—emphasis added)

The same strategy could be used by false promisers in a world with an extremely low frequency of possible suckers for false promises: just find a way to repel all but the most gullible, and then try something with the wretched ones that remain. The game of promising lives!

Korsgaard continues:

> Kant tells us that promises would be impossible if this maxim were universalized because no one would believe them. There are various ways to find a contradiction here. . . Perhaps the clearest way to bring out a logical contradiction is to say that there would be no such thing as a promise (or anyway a repayment-promise) in the world of the universalized maxim. The practice of offering and accepting promises would have died out under stress of too many violations. Thus we are imagining a world in which the agent and everyone with his purpose is making a certain sort of promise, but also a world in which there is no such thing. And this is logically inconceivable. If universalizing a maxim makes the action proposed inconceivable, then, we can get a logical contradiction. (Korsgaard 1996: 81–82)

Again, what Korsgaard presents as "the clearest way to bring out a logical contradiction" is in reality a sequence of logically dubious steps. Her claim that there would be no such thing as a promise in the world of the universalized maxim amounts to an attempted derivation of the breakdown of promising (*C*) from the universalization of the maxim of false promising (*U*), without clear awareness that the inference chain actually contains four links, each of them highly problematic.

Let us again mention the practice of one person (X) undertaking an obligation to give another person (Y) an opportunity to talk to his dead close relatives, and X receiving money from Y in exchange for that service. Now it is a universal law of nature that it is impossible to talk to dead people. And most people know this. But has this practice died out under stress of too many violations? No. Why then should we expect this to happen with the practice of promising in a world in which all people adopt the maxim of false promising?

Korsgaard briefly revisits Kant's argument about false promises in her most recent book:

> Kant's argument against the universalizability of false promising depends on the thought that in a world where people in need of money regularly offered false promises, lenders would eventually *get the idea*. They would know that these promises were insincere. (FOOTNOTE: More properly speaking, potential lenders would *always already* have got the idea. Strictly speaking, Kant's test involves imagining your maxim as a law of nature (G, 4:421) and the laws of nature are eternal, so the effects of their universalization would always already be present. (Korsgaard 2018: 129)

When Korsgaard say that “lenders would eventually get the idea” does she mean “many,” “most” or “all” lenders? She doesn’t say. To obtain a contradiction (in the will), which is the way she endorses Kant’s argument, she needs it to be “*all* lenders” but, as repeatedly argued here, this is a very implausible claim. It is a safe generalization that *no* law of nature is *known* to be true *by everyone*. Therefore, willing that it is a law of nature that everyone issues a false promise when this is in their interest would *not* entail willing that promising disappears. Since Korsgaard has to fall back on a weaker claim that many or perhaps most lenders would know the law in question, an opportunity for cheating with a false promise would still be there. So, no disappearance of promising, and no contradiction.

2.4 Encyclopaedia Britannica

In the article “Ethics” for the *Encyclopaedia Britannica* Peter Singer writes:

> One of [Kant’s] examples is as follows. Suppose that a person plans to get some money by promising to pay it back, though he has no intention of keeping his promise. The maxim of such an action might be: “Make false promises when it suits you to do so.” Could such a maxim be a universal law? Of course not. The maxim is self-defeating, because if promises were so easily broken, no one would rely on them, and the practice of making promises would cease. For this reason, the moral law would not allow one to carry out such a plan. (Singer 1998)

First, why does Singer think that the universalization of the maxim “Make false promises when it suits you to do so” would result in promises being “so easily broken” and “*no one* relying on them” (emphasis added)? Wouldn’t people realize that, typically, breaking promises given to their close friends, family members, acquaintances, neighbors and office colleagues would *not* really suit them? Consequently, wouldn’t they still keep a large proportion of their promises, which would be a basis for some degree of trust in promises?

Second, even if promises were frequently broken we would not be entitled to conclude that then *no one* would rely on them. On the contrary, almost certainly some would. (Think of those who are naïve, immature, suggestible, mentally unstable, not particularly bright, etc.)

Lastly, to reiterate, it cannot be inferred that the practice of making promises would cease (C) merely from the universal lack of trust in promises ($\sim T$).

2.5 Stanford Encyclopedia of Philosophy

> If I conceive of a world in which everyone by nature must try to deceive people any time this will get them what they want, I am conceiving of a world in which no practice of giving one’s word could ever arise and, because this is a law of nature, we can assume that it is widely known that no such practice could exist. So I am conceiving of a world in which everyone knows that no practice of giving one’s word exists. (Johnson and Cureton 2016, in the *Stanford Encyclopedia of Philosophy*)

There are two main problems here. First, it is merely asserted (rather than demonstrated) that no practice of giving one's word could ever arise in a world in which everyone by nature tries to deceive others if this would give them what they want. As shown above, it is far from clear that a convincing argument to that effect can be made.

Second, notice the sudden and unexplained transition from "widely known" (in the first sentence) to "everyone knows" (in the second sentence). It is interesting that those who defend Kant's argument are often torn between these two interpretations (see the sections on Korsgaard and Rawls). On one hand, opting for the weaker version ("In a ~*R*-world the truth of ~*R* would be *widely* known") has the advantage that it is more plausible than the stronger version ("In a ~*R*-world the truth of *R* would be *universally* known"). On the other hand, for Kant's purposes it is the stronger interpretation that is needed. Namely, the weaker version is clearly insufficient to secure his crucial conclusion ~*T* (that in a ~*R*-world *no one* will trust promises any longer). If the truth of ~*R* is *only* widely (rather than universally) known, this means that some of those ignorant of the truth of ~*R* would occasionally fall for lying promises, and ~*T* would then be false.

2.6 Roger Scruton

> [The categorical imperative] forbids the breaking of promises, for to will the universal breaking of promises is to will the abolition of promising, hence to will the abolition of the advantage that accrues to breaking promises, and so to will the abolition of my motive. Such forbidden ends of action are shown, by their confrontation with the supreme moral law, to involve the agent in a contradiction. (Scruton 2001: 86)

Already the first step in Scruton's reconstruction of Kant's reasoning (from willing the universal breaking of promises to willing the abolition of promising) is highly problematic. Although Kant's argument is supposed to start with the universalization of the maxim of false promising, Scruton begins instead with "willing the universal breaking of promises". The meaning of this phrase is far from clear. It seems to mean something like "willing that every single promise is broken", but this is much stronger than the universalization of the maxim of false promising, which is the only proper starting point of Kant's inference. Besides, understood in this way the first step would really amount to a giant and illegitimate leap toward the distant conclusion about willing the breakdown of promising (or "willing the abolition of promising," in Scruton's terminology).

Too much has been packed into that first step. As shown earlier in this paper, Kant's attempt to derive *C* (the crumbling of promising as an institution) from *U* (the universalization of the maxim of false promising) has serious problems at each of the four stages of the treacherous route from *U* to *C*. What Scruton essentially does is propose a quick shortcut whereby (a) willing *the abolition of promising* is supposed to be derived directly from (b) willing *the universal breaking of promises*. Now, if (b) is supposed to mean "willing *U*," then (b) does not entail (a). If, however, (b) is supposed to be a stronger claim than "willing *U*," then no reason is given why (b) should be accepted as a premise. All in all, a mere semblance of a proof is created but no honest toil is detectable in the effort.

Scruton offered this as the most charitable interpretation of the Kantian argument. It appears that he himself endorsed the argument because, arguably, had he seen it as fallacious he would have warned the reader about this.

2.7 John Rawls

Rawls considers the following maxim: "Try to make a deceitful promise in circumstances C (that is, when you are in embarrassing straits and need money, even though you know that you cannot repay the debt, and have no intention of doing so) in order to further your own personal advantage." Rawls then suggests that if everyone adopts that maxim, "no one can make a deceitful promise in circumstances C, as much as they would like to do so" (Rawls 2000: 170). Here is why:

> Kant assumes as a law of nature that people learn from experience and remember the past; hence once it becomes, as it were, a law of nature that everyone tries to make a false promise (in certain circumstances), the existence of the law becomes public knowledge. Everyone knows of it, and knows that others know of it, and so on. (Rawls 2000: 171)

Notice again a sudden jump from saying that something is *public knowledge* to the conclusion that it is *known by everyone*. There is no justification for this leap. "Public knowledge" refers to the knowledge that is *available to anyone*, but there is no implication that it is also *possessed by everyone.*

Even if we agree with Rawls that in his hypothetical world the universal human tendency to make deceitful promises in circumstances C would become public knowledge, this alone cannot establish his conclusion that in that world "no one could make a deceitful promise in circumstances C, as much as they would like to do so." To prove this he would need a much stronger premise, namely that this tendency is a matter of *universal* knowledge, and *not merely public* knowledge. But he gives no reason for accepting that premise. Therefore, as long as there would remain people ignorant of the universal human tendency to make deceitful promises, such promises *could* be made after all (because some people would trust them).

Speaking of the world in which it is a law of nature that everyone adopts the maxim of false promising, on what grounds does Rawls claim that "everyone knows of it, *and knows that others know of it, and so on*"? If everyone knows of law *L*, it by no means automatically follows that everyone then also knows that all others know of it, and so on.

Furthermore, a problem for Rawls would remain even if it were conceded (for the sake of argument) that in the situation he pictures *everyone* is aware of the universal human tendency to make deceitful promises. Rawls says (following Kant) that in that case all people "would laugh right away at attempts to make deceitful promises." But would they?

In Rawls's hypothetical situation all people would make deceitful promises when they needed money, but *only if* they thought that making this kind of promise would "further [their] own personal advantage" (p. 170). Presumably then, there would be situations in which people would judge that making an *honest* promise to return a loan would be to their *best* advantage (e.g. when they badly needed money but a deceitful promise to a prospective lender were regarded as too risky, dangerous or on the whole harmful). Now other people would often be unable to judge whether a person was making an honest or deceitful promise, because they would not always know enough about the promiser's circumstances. For this reason, they would have to decide whether to accept a promise, not being sure about how truthful it is. And so, because of their imperfect knowledge, people would occasionally make a mistaken decision and fall for deceitful promises. Therefore Rawls is wrong that in his imagined world people "would laugh right away at attempts to make deceitful promises".

Finally, people might even deliberately lend money to someone, *knowing* that the borrower probably has no intention to return it, simply because they would count on other means of recovering the loan than relying on the borrower's honesty. So, contrary to Kant's and Rawls's prediction that all people would laugh right away at attempts to make deceitful promises, some would actually accept them, expecting they would in the end have the last laugh.

2.8 Derek Parfit

Parfit comments on Kant's argument that the universalization of the lying promise maxim would lead to the breakdown of promising:

> In assessing this claim, as Rawls suggests, we should ask what would be true after some period that was long enough for everyone's acceptance of the lying promiser's maxim to have its full effects. Kant seems right to claim that, in such a world, no one would be able to benefit themselves by making any lying promise. Not only would such promises not be believed; the social practice of morally motivated, trust-involving promises would have ceased to exist. (Parfit 2011: 279)

Parfit agrees with Kant that in the imagined scenario "no one would be able to benefit themselves by making any lying promise" because "such promises would not be believed." Not believed *by whom*? It seems that Kant and Parfit need here the phrase "not believed *by anyone,*" because the use of any weaker expression ("not believed by many," "not believed by most" or "not believed by almost anyone") would be insufficient to establish their strong conclusion that *no one* would be able to benefit from a lying promise.

But, as pointed out earlier, the claim that lying promises would be believed by *no one* is very unlikely to be true in the world of the universalized maxim of lying promises, even "after some period that was long enough for everyone's acceptance of the lying promiser's maxim to have its full effects." Why very unlikely? Well, consider the world as it is and just imagine that a new law of nature is discovered to be true in that world. Now, whatever that law might be, what is the probability that "after some period that is long enough" *literally everyone* in that world would know for certain that this law is true? A reasonable answer is "Very close to zero."

Notice that, in contrast to Kant and many commentators who argue that the universalization of *M* would quickly result in the whole practice of promising being destroyed by too much dishonesty, Parfit is more restrained and says merely that "the social practice of *morally motivated, trust-involving* promises would have ceased to exist" (emphasis added). But Parfit's claim here is not very relevant in the context of

Kant's argument. For Kant's goal was not to find out what kind of promises continued to exist and what kind ceased to exist after the universalization of the maxim of false promises. What he really tried to prove (unsuccessfully) is that in a *U*-world it would be impossible to use any kind of promise as a way to get money by deception.

What Parfit in fact did is make Kan't argument about the breakdown of promising look more plausible by arbitrarily weakening its conclusion. While Kant wanted to show that *U* leads to the breakdown of promising, Parfit watered this down to the claim that *U* leads to the breakdown of *morally motivated, trust-involving* promising. But, of course, it is trivially true that if everyone would always break a promise whenever it would be in his *selfish* interest to do so, this would indeed be the end of *morally motivated, trust-involving* promising. Yet this truism can hardly be seen as an issue worthy of philosophical discussion.

* * *

Kant's argument about how the universalization of the maxim of false promising would lead to a contradiction is one of the best known and most widely discussed claims in the history of philosophy. Can it be that (as suggested in this article) Kant and many leading Kant scholars and ethicists failed to see several serious problems in that argument although it has been under intense scrutiny for more than 230 years? This is possible, for sure, but many will understandably think it unlikely.

Moreover, even if my diagnosis turns out to be on the right track, an explanation is still needed for why Kant's reasoning that is claimed here to be flawed on so many levels has been widely accepted for so long. Given the complexity of the issues (and lack of space) I can do no more than gesture at a potential explanation, or better at a promising explanatory factor.

At least in some cases Kant's argument about false promises may have been uncritically accepted due to excessive admiration that philosophers often show for great thinkers of the past. What points in that direction are passages involving two of the philosophers criticized here. First, John Rawls:

> I always took for granted that the writers we were studying were much smarter than I was . . . If I saw a mistake in their arguments, I supposed those writers saw it too and must have dealt with it. . . I assumed there were never plain mistakes, not ones that mattered anyway. (Rawls 2001: 427)

The problem with Rawls's attitude is that if one approaches the work of a great philosopher by *assuming* that he "never made plain mistakes, not ones that mattered anyway", then one is making it very difficult for oneself to spot such mistakes, if they exist (which is always a possibility).

Elsewhere Rawls argues similarly:

> I assume . . . that someone with Mill's enormous gifts can't be mistaken about something so basic to his whole doctrine. Little mistakes and slips, yes—they don't matter and we can fix them up. *But fundamental errors at the very bottom level: no* . . . We must have confidence in the author, especially a gifted one. If we see that something is wrong when we take the text in a certain way, then we assume the author would have seen it too. So our interpretation is likely to be wrong. We then ask: How can we read the text so as to avoid the difficulty? (Rawls 2007: 268—emphasis added)

But, if we follow Rawls's advice, how can we read the text so as to avoid a fatal difficulty that appears to present itself in a philosophical classic? One way would be to have so much confidence in the gifted author that you overestimate the strength of his argument, downplay the seriousness of the difficulty, and after failing to explore the issue fully you end up concluding, unjustifiably, that his argument does deliver the intended conclusion.

Christine Korsgaard spoke approvingly about Rawls's approach in her remarks made at a memorial service for Rawls in 2003:

> When teaching the classics of moral philosophy Jack would say: "We are not going to criticize these thinkers, but rather to interpret their positions in ways that make the best sense of them, and to see what we can learn from them." Jack had no tolerance for readers who suppose that the great thinkers of the past might be saying something completely muddled, or silly, or unintelligible. (Korsgaard 2003: 4)

In the same speech Korsgaard says that she found the model for her own work in Rawls's course on the history of ethics, the very course in which he advised students not to criticize the great thinkers. She praises Rawls for "having no tolerance for readers who suppose that the great thinkers of the past might be saying something completely muddled, or silly, or unintelligible." But isn't it, on the contrary, entirely reasonable to suppose that philosophical classics *might* be saying something completely muddled, silly, or unintelligible? After all, they were humans like us, not gods. Why should we approach their work by disallowing a possibility that they had weak moments and sometimes defended opinions that were confused, silly or even incoherent? If we make a decision that "we are not going to criticize these thinkers", how much will we be able to understand their work at all?

Adopting this kind of attitude toward Kant increases the danger that he might be "islanded by oceans of the wrong kind of respect" (Strawson 1968: 332).

Acknowledgements I am grateful to Tomislav Bracanović, Nathan Cofnas, Rafael De Clercq, Tomislav Janović, Matej Sušnik and Jared Warren for very useful comments on earlier drafts.

References

Allison, H. E. (2011). *Kant's groundwork for the metaphysics of morals: A commentary*. Oxford: Oxford University Press.

Blackburn, S. (2001). *Being good: A short introduction to ethics*. New York: Oxford University Press.

Feldman, F. (1998). Kantian ethics. In J. P. Sterba (Ed.), *Ethics: The big questions*. Malden: Blackwell.

Festinger, L., et al. (1956). *When prophecy fails*. Minneapolis: University of Minnesota Press.

Frankena, W. K. (1973). *Ethics*. Englewood Cliffs: Prentice-Hall.

Guyer, P. (2014). *Kant*. London: Routledge.

Herley, C. (2012). Why do Nigerian scammers say they are from Nigeria?. http://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/WhyFromNigeria.pdf. Accessed 1 Mar 2020.

Johnson, R., & Cureton, A. (2016). Kant's moral theory. *Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/kant-moral/. Accessed 1 Mar 2020.

Kant, I. (1997). *Groundwork of the metaphysics of morals*. Cambridge: Cambridge University Press.

Korsgaard, C. M. (1996). *Creating the kingdom of ends*. Cambridge: Cambridge University Press.

Korsgaard, C. (2003). John Rawls. *Harvard Review of Philosophy, 11*, 4–6.

Korsgaard, C. M. (2018). *Fellow creatures: Our obligations to the other animals*. Oxford: Oxford University Press.

Maclntyre, A. (1998). *A short history of ethics: A history of moral philosophy from the Homeric age to the twentieth century.* London: Routledge.

O'Neill, O. (1989). *Constructions of reason: Explorations of Kant's practical philosophy.* Cambridge: Cambridge University Press.

O'Neill, O. (1994). A simplified account of Kant’s ethics. In J. E. White (Ed.), *Contemporary moral problems*. St Paul: West Publishing Company.

Parfit, D. (2004). What we could rationally will. *The Tanner Lectures on Human Values 24*. Salt Lake City: The University of Utah Press.

Parfit, D. (2011). *On what matters (Vol. one)*. New York: Oxford University Press.

Rawls, J. (2000). *Lectures on the history of moral philosophy.* Cambridge: Harvard University Press.

Rawls, J. (2001). Afterword: A reminiscence. In J. Floyd & S. Shieh (Eds.), *Future pasts: The analytic tradition in twentieth-century philosophy.* New York: Oxford University Press.

Rawls, J. (2007). *Lectures on the history of political philosophy.* Cambridge: Harvard University Press.

Scruton, R. (2001). *Kant: A very short introduction*. Oxford: Oxford University Press.

Singer, P. (1998). Ethics. *Encyclopaedia Britannica*.

Strawson, P. F. (1968). Bennett on Kant's aanalytic. *Philosophical Review, 77*, 332–339.

Vitek, W. (1993). *Promising*. Philadelphia: Temple University Press.

Wood, A. W. (1972). Kant on false promises. In L. W. Beck (Ed.), *Proceedings of the Third International Kant Congress*. Dordrecht: Reidel.

Wood, A. W. (2008). *Kantian ethics*. Cambridge: Cambridge University Press.

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.