

# Responsibility and iterated knowledge

Alex Kaiserman

University of Oxford

## Correspondence

Alex Kaiserman, University of Oxford.

Email:

[alexander.kaiserman@balliol.ox.ac.uk](mailto:alexander.kaiserman@balliol.ox.ac.uk)

## Abstract

I defend an iterated knowledge condition on responsibility for outcomes: one is responsible for a consequence of one's action only if one was in a position to know that, for all one was in a position to know, one's action would have that consequence.

## 1 | A PUZZLE

Consider the following case:<sup>1</sup>

**LIGHT SWITCH:** D gets home from work and, as she does every evening, turns on the light switch in the hallway. Unfortunately, during the day a wire in the circuit has come loose, so that D's turning on the switch causes V, who is upstairs, to get an electric shock from his hairdryer.

D is not responsible for V's injury in **LIGHT SWITCH**, despite the fact that D acted freely in turning on the light switch and her action caused V's injury. Intuitively this is because D didn't know, nor could she reasonably have been expected to know, that her action would cause injury (to V or anyone else).

But now consider a different case:<sup>2</sup>

**ROULETTE:** D has a gun loaded with twelve indistinguishable cartridges. She knows that all but one of the cartridges are blanks, and that the cartridges are randomly distributed in the magazine. She points the gun at V and pulls the trigger. Unfortunately for V, the chamber contained the one live cartridge; a bullet is fired and V is injured.

In **ROULETTE** too, D doesn't know, nor could she reasonably have been expected to know, that her action would cause injury (to V or to anyone else). Yet D is clearly responsible for V's injury in

-----  
This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2023 The Authors. *Philosophical Issues* published by Wiley Periodicals LLC.

ROULETTE. We thus have a puzzle: what is the difference between these two cases that accounts for why D is responsible for V's injury in one but not the other?

In this paper, I argue that the relevant difference between LIGHT SWITCH and ROULETTE is a difference in what D knew, not about the consequences of her actions, but about *what she knew about* the consequences of her actions. Although D wasn't in a position to know in ROULETTE that her pulling the trigger would cause harm to V, she was in a position to know that, for all she was in a position to know, her pulling the trigger would cause harm to V; the same is not true, however, in LIGHT SWITCH. I defend this claim in section 3, before using it in section 4 to motivate a novel formulation of the epistemic condition on responsibility for outcomes.

## 2 | ALTERNATIVE EXPLANATIONS

Before outlining my preferred solution to the puzzle, I should briefly address four alternative diagnoses of the difference in our intuitions about LIGHT SWITCH and ROULETTE.

A first hypothesis is that D is responsible for V's injury in ROULETTE but not in LIGHT SWITCH because D's action was *subjectively justified* in LIGHT SWITCH but not in ROULETTE. Perhaps there were good reasons for D to turn on the switch in LIGHT SWITCH which clearly outweighed the risks, whereas there were no good reasons for D to point the gun at V and pull the trigger in ROULETTE. But this can't be the right explanation. To see this, suppose that V, in both cases, is about to culpably kill an innocent person, W. D knows this, and moreover knows that injuring V would prevent him from carrying out the killing. Now D's action is subjectively (and indeed objectively) justified in both cases. Yet our judgements about them, I submit, do not change: D is not responsible for V's injury, or for W's survival, in LIGHT SWITCH (it would be wrong of D to claim any credit for saving W given her lack of knowledge about the faulty switch); whereas D *is* responsible for V's injury, *and* for W's survival, in ROULETTE (D *can* claim credit for saving W, given that she knew there was a single live bullet in the gun).

To be clear, in saying that D is responsible for V's injury, I don't mean to imply that D is necessarily to *blame*, either for the injury or *simpliciter*; indeed D may be praiseworthy in virtue of being responsible for the injury which prevented V from killing an innocent person.<sup>3</sup> Nor do I mean to imply that D bears any forward-looking duties (for example, to compensate or apologise to V) or liabilities (for example, to be harmed in self-defence); the connections between responsibility and responsibility-*ies* is not at all straightforward.<sup>4</sup> Instead all I mean is that D is *answerable* for V's injury, in something like Smith's (2015) sense — the injury is attributable to D in a way that makes it intelligible to ask D to explain or justify her role in bringing it about, and makes her eligible for certain moral responses depending on the quality of reasons she is able to offer in support of her actions.<sup>5</sup> This concept is independent of whether D's action was justified, or the culpability or laudability of her intentions.

A second hypothesis is that D is responsible for V's injury in ROULETTE but not in LIGHT SWITCH because it was more *likely*, on D's evidence, that her action would cause injury to V in ROULETTE than it was in LIGHT SWITCH. But this can't be the right explanation either. To see this, it suffices to note that we can make the relevant probability in ROULETTE arbitrarily small simply by increasing the number of cartridges in the magazine. Suppose for example that there were a million cartridges loaded into the gun, all but one of which were blanks (and D knew all this); still, it seems to me, D would be responsible for V's injury in virtue of deciding to point the gun at V and pull the trigger, if the cartridge in the chamber turned out to be the live one.<sup>6</sup> In the

relevant sense, she ‘knew the risks’ her actions posed to V, and so is responsible for the consequences when those risks end up manifesting, no matter how small the risk.

A third hypothesis about the relevant difference is that the possibility of D’s action causing injury to V was *occurrently salient* to D in ROULETTE, but wasn’t to D in LIGHT SWITCH. Although both would presumably have agreed, if prompted, that there is a possible world in which their action harms V, we can assume that D in ROULETTE was actively thinking about that possibility when she pulled the trigger, whereas in LIGHT SWITCH that possibility presumably didn’t even occur to D. But again, this can’t be the right explanation. Suppose we ask D in LIGHT SWITCH to agree that there is a possible world in which her turning on the switch electrocutes V, thereby making her attend to that possibility; I don’t think this is enough to make D responsible for V’s injury when she turns on the switch. What D seems to lack is *knowledge* of some kind, not mere attention. Similarly, we can stipulate that D in ROULETTE *wasn’t* actively attending to the possibility where her pulling the trigger injures V — although she knew that there was a live cartridge in the magazine, the possibility of that cartridge being in the chamber simply wasn’t on her mind at the time, for whatever reason — but still, it seems to me, D is responsible for V’s injury. It will not avail her to say that she wasn’t thinking about the live cartridge when she pulled the trigger, if she knew (even if non-occurently) that it was there.<sup>7</sup>

Here is one final hypothesis to consider. In neither LIGHT SWITCH nor ROULETTE was D in a position to know that her action *would* cause injury. But in ROULETTE at least, D was in a position to know that her action *might* cause injury; and perhaps this is all that’s required for D to be responsible for V’s injury. As we’ll see, I think there is something right about this suggestion. But by itself it isn’t enough, because there is a perfectly natural sense in which D in LIGHT SWITCH is *also* in a position to know that her action ‘might’ cause injury, insofar as she is in a position to know that there is a possible world in which it does. Thus more needs to be said about the relevant meaning of ‘might’ here. (I return to this point in section 4, below.)

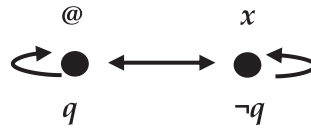
In summary, there remains a puzzle about how to explain the difference in our judgements about LIGHT SWITCH and ROULETTE. Perhaps not everyone will share the intuitions I have appealed to in this section. Or perhaps there are variants of the proposals above that avoid some or all of my criticisms of them. Nevertheless I hope at least to have motivated the need for an alternative approach. I explore such an approach in the following section.

### 3 | KNOWN UNKNOWNNS AND UNKNOWN UNKNOWNNS

Let us begin by taking a closer look at ROULETTE. D, of course, is not in a position to know that the cartridge in the chamber of her gun is a blank (since it isn’t); nor, indeed, is she in a position to know that the cartridge in the chamber is live. Moreover (and this is the important bit), exactly the same would have been true if the cartridge in the chamber *had* been a blank — D wouldn’t have been in a position to know, in such a situation, that it was a blank (or, of course, that it wasn’t). Here I am appealing to a version of what is sometimes called the ‘lottery intuition’<sup>8</sup> — if all I know is that one live cartridge has been distributed somewhere at random in a magazine of blanks, I am not in a position to know that the cartridge in the chamber isn’t the live one, however many cartridges are in the magazine. It follows that although D isn’t in a position to know in ROULETTE that her pulling the trigger would cause injury to V, she *is* in a position to know *that she isn’t in a position to know* that her pulling the trigger *won’t* cause injury to V.

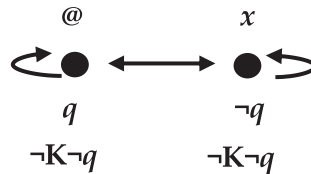
It might help to illustrate what is going on here with a bit of toy epistemic logic.<sup>9</sup> Let’s say that a possibility *v* is *epistemically accessible* from a possibility *w* iff, for all D is in a position to know

in  $w$ , she is in  $v$ . Let  $q$  be the proposition that  $D$  would cause injury to  $V$  by pulling the trigger. Finally, let's suppose for simplicity that there are just two possibilities:  $@$ , where the cartridge in the chamber is live, and  $x$ , where it's a blank. The structure of the case looks like this:



The circles and arrows represent possibilities and relations of epistemic accessibility, respectively. The relevant propositions that are true in each possibility are labelled underneath.

It's easy to see that  $D$  isn't in a position to know that  $\neg q$  either in  $@$  or in  $x$ , since there is a world accessible from both  $@$  and  $x$  in which  $q$  is true (namely,  $@$ ). Using 'Kp' as a shorthand for ' $D$  is in a position to know that  $p$ ', we can thus update our diagram as follows:



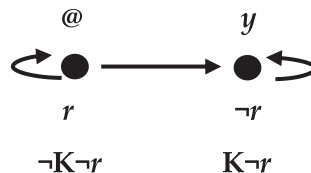
It's now easy to see that  $D$  is in a position to know that *she isn't in a position to know* that  $\neg q$  (i.e.  $K\neg K\neg q$ ) in  $@$ , since  $\neg K\neg q$  is true in every possibility accessible from  $@$ .

Next let's look at LIGHT SWITCH. In fact, the light switch is defective.  $D$  however isn't in a position to know that the light switch is defective, nor (of course) that it isn't. But (and this is the important bit) if the switch *hadn't* been defective,  $D$  *would* have been in a position to know that it wasn't. Here I am appealing to a broadly anti-sceptical assumption, namely that in 'normal' possibilities, we are in a position to know that 'abnormal' possibilities don't obtain. I know that I have hands, for example, even though there's a possible world that is subjectively indistinguishable from the actual world in which I don't have hands (one in which I'm a brain in a vat, say). It follows that not only is  $D$  not in a position to know that her turning on the light won't cause harm to  $V$ , she also isn't in a position to know that *she isn't in a position to know* that her turning on the light won't cause harm to  $V$ .

Here again is a toy model to illustrate this point. Let  $r$  be the proposition that  $D$  would cause injury to  $V$  by turning on the light, and as before, let's suppose for simplicity that there are just two possibilities:  $@$  (where the switch is faulty) and  $y$  (where it isn't). I claim that the epistemic structure of this case looks like this:



We can see that while  $D$  isn't in a position to know either that  $r$  or that  $\neg r$  in  $@$ ,  $D$  is in a position to know that  $\neg r$  in  $y$  (since in every world accessible from  $y$  — i.e.,  $y$  itself —  $\neg r$  is true). Hence we have:



It's now easy to see that D isn't in a position to know *that she isn't in a position to know* that  $\neg r$  (i.e.  $\neg K\neg K\neg r$ ) in @, since there is a possible world accessible from @, namely *y*, in which she is in a position to know that  $\neg r$ .

Of course, not everyone will agree with the claims I have made about what D is in a position to know in these cases. Sceptics will insist that one is never in a position to know that one's flipping a light switch won't electrocute someone, if there are possibilities subjectively indistinguishable from actuality in which the switch is faulty. Others might insist that, insofar as D in LIGHT SWITCH is in a position to know that the light switch isn't faulty when it isn't, D in ROULETTE is also in a position to know that the cartridge in the chamber of her gun is a blank when it is. But I take it there is an important difference in our judgements about these two cases. Moreover, that difference is naturally accommodated by recent *normality* accounts of knowledge, according to which (to a first approximation) the possibilities epistemically accessible from *w* are those which are *at least as normal as w*.<sup>10</sup> In cases like ROULETTE, D doesn't know which possibility she's in because both are equally normal (as Smith (2016) notes, no explanation would be required if the cartridge in the chamber turned out to be live, no matter how unlikely that possibility). But unlike similarity, comparative normality is not symmetric<sup>11</sup> — the possibility where the switch is faulty is less normal than the possibility where it isn't (while no explanation would be required if the switch was working properly, we would expect an explanation if the switch turned out to be faulty). This means that in cases like LIGHT SWITCH, D can know she's in the normal possibility when she's in the normal possibility, but can't know she's in the abnormal possibility when she's in the abnormal possibility (even though the possibilities are subjectively indistinguishable).<sup>12</sup> To be clear, I don't necessarily wish to endorse these normality approaches to knowledge. Nor, indeed, am I suggesting that the relevant notion of 'normality' can ultimately be understood independently of the concept of knowledge it is being used to characterise.<sup>13</sup> My point is merely that there are solid theoretical reasons to think that the epistemic structures of ROULETTE and LIGHT SWITCH do differ in the ways I have described.

So far all I have done is argue that there is a difference in what D knows about what she knows in ROULETTE and LIGHT SWITCH. But my hypothesis, of course, is that this difference also explains the difference in our judgements about responsibility. Defending this hypothesis will be the goal of the following section.

## 4 | AN ITERATED KNOWLEDGE CONDITION ON RESPONSIBILITY

Consider the following condition on responsibility for outcomes:

Iterated Knowledge Condition (IKC): S is responsible for *o* in virtue of  $\varphi$ -ing at *t* only if S was in a position to know at *t* that, for all she was in a position to know at *t*, her  $\varphi$ -ing would cause *o* (or something like it).

(IKC) is deliberately vague, in two respects. Firstly, there is a debate in the literature about whether the epistemic condition on responsibility should be formulated in terms of S's causing *o* or in terms of S's causing an outcome of a certain *kind*, of which *o* was an instance;<sup>14</sup> I wish to remain neutral on this question, hence the parenthetical qualification. Secondly, there are various ways of interpreting the phrase 'in a position to know that *p*'. One option is to interpret it narrowly, as expressing something like the idea that S would have known that *p* had she believed that *p* (and her epistemic situation had otherwise remained the same). Another option is to interpret it relative to

some objective normative standard: S is in a position to know that  $p$  just in case a *reasonable person* in S's position would have known that  $p$ . Which option one prefers will inevitably be tied up with one's views about the success of the 'tracing strategy', which attempts to trace all responsibility for unwitting wrongdoing back to some earlier action performed in full knowledge of the risks.<sup>15</sup> Much has been written on these questions, but again I wish to remain neutral with respect to them; one should thus feel free to interpret the qualifier 'in a position to' as expressing whatever modality one wishes. Finally, note that (IKC) is a condition on S being responsible for  $o$  in virtue of  $\varphi$ -ing at  $t$ ; as is well known, an agent may causally contribute to an outcome via multiple actions or omissions at different times, and may satisfy the epistemic condition on responsibility with respect to some of these actions or omissions but not others.<sup>16</sup> So long as there is *some*  $\varphi$  and  $t$  such that S is responsible for  $o$  in virtue of  $\varphi$ -ing at  $t$ , she is responsible for  $o$  *simpliciter*.

Though (IKC) may appear complex, there are more familiar (albeit more ambiguous) ways of expressing the same idea. One makes use of the epistemic 'might'. Assuming that there is an interpretation of 'It might be the case that  $p$ ' on which it is equivalent to 'For all S is in a position to know at  $t$ ,  $p$ ',<sup>17</sup> (IKC) can be stated as follows:

(IKC\*): S is responsible for  $o$  in virtue of  $\varphi$ -ing at  $t$  only if S was in a position to know at  $t$  that her  $\varphi$ -ing *might* cause  $o$  (or something like it).

Several authors endorse claims very similar to (IKC\*), usually once they realise that the condition obtained by replacing 'might' in (IKC\*) with 'would' is much too strong.<sup>18</sup> The problem of course is that there are many non-epistemic interpretations of 'might', and on many of these (IKC\*) is obviously far too weak. As pointed out in section 2, above, there is a perfectly natural sense in which D was in a position to know that her turning on the light 'might' cause harm to V in LIGHT SWITCH, insofar as she was in a position to know that there's a possible world in which this happens.<sup>19</sup> (IKC) can thus explain the appeal of (IKC\*), while at the same time avoiding ambiguity by stating the condition explicitly in terms of iterated knowledge.<sup>20</sup>

(IKC) is unusual among statements of the epistemic condition on responsibility in the literature in not making use of the concept of 'reasonable foreseeability'.<sup>21</sup> This is deliberate. 'Foresee', like other similar verbs (e.g. 'see', 'remember', 'hear', etc.), can take either a noun phrase or a that-clause as its object. This gives rise to two different versions of the reasonable foreseeability condition, which aren't always cleanly distinguished in the literature:<sup>22</sup>

(RF-1): S is responsible for  $o$  in virtue of  $\varphi$ -ing at  $t$  only if  $o$  was reasonably foreseeable by S at  $t$ .

(RF-2): S is responsible for  $o$  in virtue of  $\varphi$ -ing at  $t$  only if it was reasonably foreseeable by S at  $t$  that her  $\varphi$ -ing might cause  $o$  (or something like it).

(RF-1) is both too weak and too strong.<sup>23</sup> To see why it's too weak, consider the following case:

VOTER: As an informed insider, D foresaw the Red Team's election victory. But he decides to vote anyway. Unfortunately, the voting machine he used was labelled wrong — pressing the button labelled 'Blue Team' registers a vote for the Red Team and vice versa. D presses the button labelled 'Blue Team', intending to vote for the Blue Team.



D is not responsible for the Red Team's election victory in VOTER, even though he foresaw it, given that he had no reason to think his actions would or might causally contribute to the victory.

To see why (RF-1) is too strong, consider the following case:

SHOOTER: D fires a gun several times into a densely packed crowd of people. Several people are badly injured.

D is clearly responsible for the injuries caused by her actions in SHOOTER. Yet none of the injuries that actually occurred were reasonably foreseeable by her; after all, D couldn't have known who her bullets would hit, and so wasn't in a position to discriminate any particular future injury from any of the others.<sup>24</sup>

(RF-2), by contrast, gets the right results in these cases. But now a further issue arises. Foreseeing that *p* is what is known as a *factive mental state* — you can't foresee that *p* unless *p* is true. Williamson (2000, Ch.1) influentially argues that *knowledge* is the most general factive mental state. This means that foreseeing that *p*, just like seeing, hearing, or remembering that *p*, is a particular way of knowing that *p*. Seen in this light, though, it's hard to justify (RF-2)'s focus on *foresight* in particular over all the other ways of coming to know about the future. Suppose someone tells me that if I press a certain button, a bomb will explode; or I see a sign warning me of this; or it was me who planted the bomb and connected it to this button in the first place. In none of these cases do I *foresee* that my pressing the button will cause an explosion.<sup>25</sup> Nevertheless I know that it will, by some means or other — I *hear* that it will, or *see* that it will, or *remember* that it will — and this seems clearly enough to make me responsible for the explosion if I press the button. This appears to speak in favour of replacing 'foresee' in (RF-2) with 'know'. But then what we get is just (IKC\*), which on its most plausible interpretation (or so I have argued) is equivalent to (IKC).

Another way of thinking of (IKC) is as an elucidation of the familiar (Rumsfeldian) idea that one is responsible for the *known unknown* consequences of one's actions, but not for the *unknown unknown* ones.<sup>26</sup> Very often we don't know what the consequences of our actions will be. Nevertheless, when we *know* (or at least are in a position to know) that we are ignorant about whether our actions will have a certain consequence, we must take that risk into account in our decision-making. We may decide to take the risk, and indeed may be justified in doing so; but we are answerable for the consequences if that risk ends up manifesting. By contrast, we cannot be responsible for something when we couldn't have known *even that we were ignorant* about whether it would happen. What the arguments of section 3 show is that we can fail to know that for all we know something will happen even if the possibility of its happening is one we have explicitly considered. In LIGHT SWITCH, for example, D is not even in a position to know that, for all she's in a position to know, turning on the switch would electrocute V — not because she hasn't considered that possibility, but because for all she is in a position to know, the light switch is working normally, and if it *had* been working normally she *would* have known that it was.

In summary, (IKC) is an independently plausible condition on responsibility, which can be motivated from several different directions. But the primary motivation for (IKC), of course, is that it solves the puzzle with which we began. It is compatible with (IKC) that D is responsible for V's injury in ROULETTE, because D is in a position to know that she isn't in a position to know that her pulling the trigger won't harm V; but D is not responsible for V's injury in LIGHT SWITCH, according to (IKC), because for all D is in a position to know, she *is* in a position to know that her turning on the light switch won't harm V. (IKC) can thus explain the difference in our judgements about the two cases. Moreover, (IKC) avoids the problems with the alternative

hypotheses considered in section 2, since the verdict delivered by (IKC) is entirely insensitive to the probability on D's evidence of her action causing harm to V, whether D's action was subjectively or objectively justified, or whether the possibility of D's action causing harm was occurrently salient to her.

That said, (IKC) also makes concrete predictions about the sorts of changes to LIGHT SWITCH and ROULETTE which *would* change our judgements about them. As discussed, increasing the number of blanks in D's gun in ROULETTE intuitively makes no difference to her degree of responsibility for V's injury, so long as she knew that it contained one live bullet. But now suppose D's gun was *supposed* (in some sense of the word) to contain only blanks — perhaps D is an actor, and the guns on set are supposed to be prop guns. Something has gone wrong, however, and the gun that has ended up in D's hand contains live bullets. D *could* have opened up the gun to check its contents; if prompted, she would have agreed that there is a possible world in which her gun contains live bullets; we can even stipulate that D had no good reason to pull the trigger (she was just bored). Nevertheless, (IKC) predicts that D isn't responsible for V's injury in this case, since for all she's in a position to know the gun *is* a prop gun, and *had* it been a prop gun she would have been in a position to know that it was. This strikes me as the right result.<sup>27</sup>

Or consider LIGHT SWITCH again. As discussed, it's not enough for D to be responsible for V's injury in LIGHT SWITCH that D knows there is a possible world in which the switch is faulty. But now suppose D has some further evidence — for example, she knows that a small fraction of every light switch of the kind she owns has a manufacturing fault which makes it prone to fail in ways that cause electrocution. Then she would be responsible for V's injury, according to (IKC), since she now knows that she *doesn't* know that her action *won't* cause injury to V. This again strikes me as the right result.

It might be objected that (IKC) is too strong in requiring *knowledge*, rather than mere belief. Consider LIGHT SWITCH again, but this time let's suppose that a qualified electrician has told D that the light switch is defective. In fact, the electrician never checked the switch — perhaps she was on another job, but doesn't want to admit this to D. If D believes that the light switch is defective on the electrician's say so, her belief is true and justified, but it doesn't amount to knowledge, since it is only a matter of luck that her belief is true (this is a Gettier-style case). But now suppose D decides to flip the switch, hoping and intending to electrocute V by doing so, which is exactly what happens. According to (IKC), D is not responsible for V's injury. Doesn't this seem like the wrong result?

I think we need to be careful here to tease apart a few different judgements. Obviously D is blameworthy for flipping the switch. She might even be liable to pay compensation to D — as Tadros (2021) argues, being culpable with respect to an outcome can be an independent ground for liability, even if one isn't responsible for the outcome. Nevertheless I want to insist that D is not responsible for V's injury in this case. Of course this means that whether one satisfies the epistemic condition on responsibility is partly outside of one's control. But this is true of the other conditions on responsibility for outcomes too — it is a necessary condition on being responsible for an outcome that one's actions *caused* the outcome, for example, even though it is partly outside of one's control whether one satisfies this condition.

The point might be easier to see if we consider a case with the opposite valence. Suppose an electrician has told D that flipping the blue switch turns off the power. D sees V trying to change a light fitting, and quickly flips the blue switch hoping to prevent V from electrocuting himself. In fact this does turn off the power, but it turns out the electrician didn't know that the blue switch is the power switch, and so neither did D. Once all this is revealed to D, I think it would be wrong of



her to claim credit for saving V. Of course her intentions were laudable, and her actions did cause V's survival, but it is only a matter of luck that her actions had the desired effect.

To be clear, what I am saying here is compatible with the view (endorsed by Rosen (2008, p. 596), Peels (2014, pp. 493–4), Baron (2017, pp. 58–9) and others) that whether or not a particular belief amounts to knowledge makes no difference to one's degree of blameworthiness or praiseworthiness. But the topic of this paper is responsibility for outcomes, not blameworthiness and praiseworthiness; and I do think there is something importantly right about the idea that responsibility for outcomes requires *knowledge* of some kind, or at least the possibility of such knowledge; not mere belief. The challenge is to figure out *what* exactly must be known in order to be responsible for some outcome; this is the question I have sought to answer above. And as it turned out, my explanation of the difference between LIGHT SWITCH and ROULETTE relied crucially on structural properties of knowledge in particular, and couldn't have been replicated using a condition that refers only to what the agent believes.

## 5 | CONCLUSION

I began this paper with a puzzle: what explains the differences in our judgements about responsibility in cases like LIGHT SWITCH and cases like ROULETTE, given that in neither case is D in a position to know that her action would cause injury? I argued that the right explanation can't be anything to do with whether D's action was justified, the probability of the outcome occurring on D's evidence, or to what extent D was attending to the possibility of her action having that consequence. Instead, I argued that the relevant difference has to do with what D was in a position to know *about what she was in a position to know* about the consequences of her action. I used this insight to motivate a novel iterated knowledge condition on responsibility for outcomes, which can explain the appeal of rival statements of the epistemic condition, while also avoiding their shortcomings.

Looking back at LIGHT SWITCH and ROULETTE with the benefit of hindsight, it's easy to see the appeal of something like (IKC). Much contemporary philosophical work on knowledge in recent years can be seen as concerned, more or less directly, with resolving a tension between two conflicting intuitions: the 'lottery intuition', that one is not in a position to know that one's lottery ticket is a loser no matter how likely that is, and the 'Moorean intuition' that in normal cases we know lots of things, even things that are false in possibilities subjectively indistinguishable from the actual world. In retrospect it is striking how closely LIGHT SWITCH and ROULETTE resemble the sorts of cases used to illustrate the contrast between lottery cases and non-lottery cases in epistemology. It would be no surprise, then, if the two puzzles turned out to have a common solution.

## ACKNOWLEDGEMENTS

Thanks to Daniel Kodsi, Kida Lin, Daniel Miller, Ana Radomirescu, Carolina Sartorio, Philip Swenson, Elad Uzan, and audiences in Berlin, Bratislava, Oxford, London and Sussex for helpful feedback on previous versions of this paper.

## ENDNOTES

<sup>1</sup> Adapted from one in Thomson (1990, p. 229).

<sup>2</sup> Also adapted from one in Thomson (1986, p. 181).

<sup>3</sup> See also Talbert (2023, pp. 432–4) on this point.

<sup>4</sup>See Tadros (2021) on the connections between responsibility, liability and culpability.

<sup>5</sup>This is intended as a functional characterisation of responsibility – an elucidation of the role responsibility plays in our ethical practices – rather than an analysis of it. It is an open question whether there is a single relation that plays the role described above (Smith (2012) thinks there is; Shoemaker (2011) disagrees). Nevertheless I will be assuming for the purposes of this paper that there is such a relation, and that we have a good enough pre-theoretic grasp of it to permit theoretical investigation into the conditions under which it holds.

<sup>6</sup>Interestingly, this point is clearly anticipated by Mason J. in *Wyong Shire Council v Shirt* (1980) 146 C.L.R. 40, a seminal Australian negligence case: “[W]hen we speak of a risk of injury as being ‘foreseeable’ we are not making any statement as to the probability or improbability of its occurrence save that we are implicitly asserting that the risk is not one that is far-fetched or fanciful”. This passage has been widely criticized for reducing the reasonable foreseeability requirement on liability to triviality (see, e.g., *Tame v New South Wales* [2002] HCA 35). I think the critics are mistaken, for reasons ably revealed by cases like ROULETTE – the possibility of the gun going off in ROULETTE certainly doesn’t seem ‘farfetched or fanciful’, no matter how many blanks are in the magazine. ‘Farfetched and fanciful’ is better contrasted with *normal* than it is with *probable*. See section 3, below.

<sup>7</sup>Something like this point has been made in the context of the debate over how to distinguish between the legal concepts of recklessness and negligence – see, e.g., Husak (2011, pp. 208–9); c.f. Zimmerman (2017, p. 79).

<sup>8</sup>See, e.g., Hawthorne (2003); Pritchard (2005).

<sup>9</sup>I should stress that these are *toy* models. In particular, any attempt to model ‘in a position to know that’ as a necessity modal involves accepting certain idealisations, for example that one is in a position to know every necessary truth. Although some epistemologists (notably Williamson (2000)) do appear to assume that ‘in a position to know that’ has a normal modal logic, Yli-Vakkuri and Hawthorne (2022) point out that this leads to implausible consequences when combined with the claims that knowing entails being in a position to know and being in a position to know entails possibly knowing. In any case, it is certainly no part of my use of the phrase ‘in a position to know’ that it has a normal modal logic, even if modelling it as if it does can be helpful for visualising the differences in epistemic structure between LIGHT SWITCH and ROULETTE.

<sup>10</sup>See, e.g., Stalnaker (2006), Greco (2014), Loets (2022), Goodman and Salow (2023).

<sup>11</sup>As Goodman and Salow (2023) note, it’s this difference which crucially distinguishes normality theories from safety- or sensitivity-based theories.

<sup>12</sup>c.f. Magidor (2018), who argues using a broadly safety-based approach to knowledge that both me and my brain-in-a-vat counterpart are in a position to know whether we’re envatted. Like Goodman and Salow (2023), I’m inclined to think of this as a *reductio* of the safety approach.

<sup>13</sup>See Goodman and Salow (2023), who note that the best versions of safety/sensitivity views must similarly reject the possibility of giving an account of the relevant notion of ‘closeness’ between worlds which does not itself presuppose epistemic notions.

<sup>14</sup>See, e.g., Fischer and Tognazzini (2009), Miller (2017).

<sup>15</sup>For critical discussion of the tracing strategy, see especially Sher (2009) and Vargas (2005).

<sup>16</sup>See, e.g., Fischer and Tognazzini (2011).

<sup>17</sup>This is controversial; for discussion, see Egan and Weatherston (2011).

<sup>18</sup>Here’s a representative passage, from Ginet (2000, pp. 269–70, emphasis in original): “Should we say, then, that...S is blameworthy for bringing about harm H by making movement M at t1 (or for failing to prevent harm H by making movement M at t1) only if it is also the case that at t1 S knew that making movement M would bring about H (or knew that making M would prevent H)? No, this would be to require S to know too much...[W]e should not require for blameworthiness more than that at t1 S knew that making movement M would *or might* cause H.”

<sup>19</sup>Similar considerations apply to the idea that one is responsible for an outcome only if one ‘knew the risks’ of one’s actions. As Hansson (2021) notes, there is a use of the word ‘risk’ on which “when there is a risk, there must be something that is unknown or has an unknown outcome”, and hence, on this interpretation, “knowledge about risk is knowledge about lack of knowledge”. But as Hansson acknowledges, there are several other uses of ‘risk’ as well. For example, on one natural interpretation, ‘knowing the risk’ of one’s action requires knowing the precise probability of one’s action causing that outcome, and this is clearly too strong a condition on responsibility (D is still responsible for V’s injury in ROULETTE, for example, even if she didn’t know exactly how many blanks were in the magazine, so long as she knew there was at least one live bullet).

- <sup>20</sup>The only other attempt of which I am aware to appeal to iterated knowledge in stating the epistemic requirements on responsibility is Guerrero (2007), who endorses a principle he calls 'Don't Know Don't Kill': "If someone knows that she doesn't know whether a living organism has significant moral status or not, it is morally blameworthy for her to kill that organism" (Guerrero, 2007, pp. 78–9). Guerrero's aims in that paper are quite different from mine, however; in particular, Guerrero is not concerned with solving the sort of puzzle I described in section 1.
- <sup>21</sup>The concept of reasonable foreseeability originates in Anglo-American negligence law, where it plays an important role in determining both whether a defendant bore a duty to the claimant not to cause them a certain kind of harm and which consequences of a breach of duty a defendant can be found liable for. For a helpful introduction to the history of the concept, and in particular the historical antecedents to Lord Atkins' famous 'neighbour principle' introduced in *Donoghue v Stevenson* ([1932] AC 562), see Scott (2019).
- <sup>22</sup>As a representative example, see Vargas (2005), who starts with the view that "[f]or an agent to be responsible for some outcome... the outcome must be reasonably foreseeable for that agent at some suitable prior time" – i.e. (RF-1) – but later equates this with the view that the agent must "reasonably foresee that his resultant behaviour might have [certain consequences]" – i.e. (RF-2).
- <sup>23</sup>Here I draw on Kaiserman (2017, §III).
- <sup>24</sup>c.f. Dretske (1969, pp. 18–35), who influentially argues that seeing an *x*, though it doesn't require knowledge or belief that it is an *x*, nevertheless requires at least the ability to visually differentiate the *x* from its immediate surroundings (see also Siegel (2006, p. 434)). I can see the Mazda 2 parked outside because I can visually differentiate it from the other cars, the curb, and the road, wall, trees and sky in the background, even if I have no idea what make of car it is (or even that it's a car). By contrast, I can look straight at a chameleon without actually seeing it, if it is so well camouflaged that I am not able to visually differentiate it from its surroundings. Of course, it's less clear what such a capacity of differentiation amounts to in the case of foreseeing than it is in the case of seeing. Nevertheless, it seems plausible that foreseeing *x* similarly requires that one is able to differentiate *x*, via whatever capacity is involved in foresight, from other future events.
- <sup>25</sup>Foresight is the future-directed manifestation of what psychologists call 'mental time travel', the ability to project oneself into different moments in time; see, e.g., Suddendorf and Corballis (2007).
- <sup>26</sup>This distinction was of course thrust into the mainstream after Donald Rumsfeld's notorious Department of Defence news briefing in 2002, but it was apparently common parlance in project management and strategic planning circles before then (see, e.g., Courtney, Kirkland and Viguerie (1997)).
- <sup>27</sup>This example bears some similarities to the events leading up to the tragic death of Halyna Hutchins on the set of *Rust* in October 2021. Interestingly, though, the charges against Alec Baldwin in that case were allegedly dropped only when evidence emerged that the gun had been modified in a way that made it possible for it to fire without the trigger being pulled.

## REFERENCES

- Baron, M. (2017). Justification, Excuse, and the Exculpatory Power of Ignorance. In R. Peels (ed.), *Perspectives on Ignorance from Moral and Social Philosophy*. Routledge.
- Courtney, H., Kirkland, J., & Viguerie, P. (1997). Strategy Under Uncertainty. *Harvard Business Review*, November – December 1997.
- Dretske, F. (1969). *Seeing and Knowing*. University of Chicago Press.
- Egan, A., & Weatherson, B. (eds.) (2011). *Epistemic Modality*. Oxford University Press.
- Fischer, J. M., & Tognazzini, N. A. (2009). The Truth About Tracing. *Noûs*, 43(3), 531–556.
- Ginet, C. (2000). The Epistemic Requirements for Moral Responsibility. *Noûs*, 34(s14), 267–277.
- Goodman, J., & Salow, B. (2023). Epistemology Normalised. *Philosophical Review*, 132(1), 89–145.
- Greco, D. (2014). Could KK Be OK? *Journal of Philosophy*, 111(4), 169–197.
- Guerrero, A. A. (2007). Don't Know, Don't Kill: Moral Ignorance, Culpability, and Caution. *Philosophical Studies*, 136(1), 59–97.
- Hansson, S. O. (2021). Risk. In E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2021 Edition), <https://plato.stanford.edu/archives/win2021/entries/risk/>
- Hawthorne, J. (2003). *Knowledge and Lotteries*. Oxford University Press.

- Husak, D. (2011). Negligence, Belief, Blame, and Criminal Liability: The Special Case of Forgetting. *Criminal Law and Philosophy*, 5(2), 199–218.
- Kaiserman, A. (2017). Partial Liability. *Legal Theory*, 23(1), 1–26.
- Loets, A. J. (2022). Choice Points for a Theory of Normality. *Mind*, 131(521), 159–191.
- Magidor, O. (2018). How Both You and the Brain in a Vat Can Know Whether or Not You Are Envatted. *Aristotelian Society Supplementary Volume*, 92(1), 151–181.
- Miller, D. J. (2017). Reasonable Foreseeability and Blameless Ignorance. *Philosophical Studies*, 174(6), 1561–1581.
- Peels, R. (2014). What Kind of Ignorance Excuses? Two Neglected Issues. *Philosophical Quarterly*, 64(256), 478–496.
- Pritchard, D. (2005). *Epistemic Luck*. Oxford University Press.
- Rosen, G. (2008). Kleinbart the Oblivious and Other Tales of Ignorance and Responsibility. *Journal of Philosophy*, 105(10), 591–610.
- Scott, H. (2019). The History of Foreseeability. *Current Legal Problems*, 72(1), 287–314.
- Sher, G. (2009). *Who Knew?: Responsibility Without Awareness*. Oxford University Press.
- Shoemaker, D. (2011). Attributability, Answerability, and Accountability: Toward a Wider Theory of Moral Responsibility. *Ethics*, 121(3), 602–632.
- Siegel, S. (2006). How Does Visual Phenomenology Constrain Object-Seeing? *Australasian Journal of Philosophy*, 84(3), 429–441.
- Smith, A. M. (2012). Attributability, Answerability, and Accountability: In Defense of a Unified Account. *Ethics*, 122(3), 575–589.
- Smith, A. M. (2015). Responsibility as Answerability. *Inquiry: An Interdisciplinary Journal of Philosophy*, 58(2), 99–126.
- Smith, M. (2016). *Between Probability and Certainty: What Justifies Belief*. Oxford University Press.
- Stalnaker, R. (2006). On Logics of Knowledge and Belief. *Philosophical Studies*, 128(1), 169–199.
- Suddendorf, T., & Corballis, M. C. (2007). The Evolution of Foresight: What is Mental Time Travel, and is it Unique to Humans? *Behavioural and Brain Sciences*, 30(3), 299–313.
- Tadros, V. (2021). Two Grounds of Liability. *Philosophical Studies*, 178(11), 3503–3522.
- Talbert, M. (2023). Causal Involvement, Collectives and Blame: Replies to Petersson. In A. Garcia, M. Gunnemyr and J. Werkmäster (eds.), *Value, Morality & Social Reality: Essays dedicated to Dan Egonsson, Björn Petersson & Toni Rønnow-Rasmussen*, Lund: Department of Philosophy, Lund University.
- Thomson, J. J. (1986). *Rights, Restitution, and Risk*. Harvard University Press.
- Thomson, J. J. (1990). *The Realm of Rights*. Harvard University Press.
- Vargas, M. (2005). The Trouble with Tracing. *Midwest Studies in Philosophy*, 29(1), 269–290.
- Williamson, T. (2000). *Knowledge and its Limits*. Oxford University Press.
- Yli-Vakkuri, J., & Hawthorne, J. (2022). Being in a Position to Know. *Philosophical Studies*, 179(4), 1323–1339.
- Zimmerman, M. (2017). Ignorance as a Moral Excuse. In R. Peels (ed.), *Perspectives on Ignorance from Moral and Social Philosophy*. Routledge.

**How to cite this article:** Kaiserman, A. (2023). Responsibility and Iterated Knowledge. *Philosophical Issues*, 33, 83–94. <https://doi.org/10.1111/phils.12244>