# Father, don't forgive them, for they could have known what they're doing

**Lara Kirfel** (l.kirfel@stanford.edu), Department of Psychology, Stanford University

**Xenia Bunk** (xenia.bunk@gmx.de), Department of Psychology, Ludwig Maximilians University of Munich

**Ro'i Zultan** (zultan@bgu.ac.il), Department of Economics, Ben Gurion University

**Tobias Gerstenberg** (gerstenberg@stanford.edu), Department of Psychology, Stanford University

## Abstract

What someone knew matters for how we hold them responsible. In three studies, we explore people's responsibility judgments for negative outcomes to knowledgeable versus ignorant agents. We manipulate whether agents arrived at their knowledge state unintentionally or willfully. In Experiment 1, agents who knew about the harmful consequences of their actions were judged highly responsible no matter how they came to know. In contrast, willfully ignorant agents were judged more responsible than unintentionally ignorant agents. Participants inferred that willfully ignorant agents were more likely to believe that their action might cause harm. When we explicitly stipulate the agents' beliefs in Experiment 2, the 'willful ignorance' effect reduces but persists. Participants inferred that the willfully ignorant agent was more likely to have acted anyhow even if they had known. Explicitly stating whether the agent's action depended on their knowledge further reduced the 'willful ignorance' effect in Experiment 3.

**Keywords:** responsibility; moral judgment; ignorance; willful ignorance; epistemic states; inferences

## Introduction

Being morally responsible requires being aware of certain things. For an agent to be held morally responsible for a negative outcome, it is usually assumed that the agent causally contributed to the outcome, and that the agent was aware of the moral consequences of their action (Wieland, 2017). Even though an agent's epistemic state is central to whether we hold them responsible (Cushman, 2008; Gerstenberg & Lagnado, 2012; Lagnado & Channon, 2008; Young & Tsoi, 2013), psychologists have only just begun to study how exactly ignorance affects judgments of responsibility and blame(Hertwig & Engel, 2016; Sargent & Newman, 2021). Ignorance and knowledge are often thought of as opposites, with knowledge being the preferred mental state because being knowledgeable usually means being in a better position to make good decisions (Nicolas, 2004). Especially when it takes effort to acquire knowledge, it's expected that this knowledge is put to good use (Kirfel & Lagnado, 2021). So when an agent's actions result in harmful consequences, knowledgeable agents are usually blamed more than ignorant ones (Cushman, 2008; Gerstenberg & Lagnado, 2012; Lagnado & Channon, 2008; Young & Tsoi, 2013). But does it only matter whether or not an agent knows, or does it also matter *how* their epistemic state came about? Imagine that the latest product a CEO has decided to launch was lucrative but had detrimental environmental consequences. The CEO was unaware that environmental harm would result from the launch. How do we evaluate the CEO's responsibility upon learning that, although ignorant, they did have access to a research report on the product's environmental effects that they deliberately decided not to look into?

### What the eyes don't see ...

People sometimes actively avoid learning information about the consequences of their behavior – they choose to remain "wilfully ignorant" (Sarch, 2018). One reason to remain ignorant is to pursue selfish interests, and to feel less bad about doing so (Conrads & Irlenbusch, 2013; Dana, 2006; Dana, Weber, & Kuang, 2007). In dictator games, where the dictator chooses how a sum of money is split between them and the recipient, dictators often choose not to know how much the recipient will receive (Nyborg, 2011; Shepperd & Howell, 2015). Avoiding knowing whether one's action hurts others provides a "moral wiggle room" to eschew responsibility (Conrads & Irlenbusch, 2013).

The case of "willful ignorance" raises the question of whether all kinds of ignorance reduce responsibility for negative outcomes, or whether it matters how the agent's ignorance came about (Kirfel & Hannikainen, 2022; Kirfel & Lagnado, 2021). Are agents blameworthy for strategically ignorant behavior (Wieland, 2017)? If so, how does the deliberate avoidance of knowledge fare against other epistemic states? Is being willfully ignorant worse than being unintentionally ignorant? Or is acting out of deliberate ignorance even worse than acting knowingly? In a series of experiments, we investigate how people hold willfully ignorant agents responsible for negative outcomes, and how this compares to judgments about knowing, or unintentionally ignorant agents.

### Integrating epistemic states into responsibility

We contrast three hypotheses about the way in which ignorance may affect judgments of responsibility. Some have argued that ignorance should *always* count as exculpatory (Phillips, 2019; Ross, 2011; Smith, 1983) (**Hypothesis #1:** *"Blameless Ignorance"*). An agent's ignorance – however it came about – fully excuses them for performing an unwitting but wrongful act. While an agent might be blamed for how they became ignorant in the first place, their acting out of ignorance should not be evaluated differently than that of some-



In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society.* 

one whose ignorance was unintentional: the ignorance of the consequences of their action equally exculpates them for any negative outcome that might arise from it (Ross, 2011).

Others have argued that an agent's responsibility is influenced by their degree of belief (Buchak, 2014; Chockler & Halpern, 2004; Edwards, 1954). Often there is uncertainty about how much an agent knew about the consequences of their action. In such cases, we need to infer others' beliefs from their actions (Aboody, Davis, Dunham, & Jara-Ettinger, 2021; Gerstenberg et al., 2018; Langenhoff, Wiegmann, Halpern, Tenenbaum, & Gerstenberg, 2021). If an agent deliberately refrains from finding out what could happen, this suggests that they have a hunch that it might be affecting others negatively (see e.g. Gorman, 2011; Husak & Callender, 2019). **Hypothesis #2:** *"Epistemic Inferences"* posits that we hold others responsible to the extent that we believe that they knew about the negative consequences of their action.

Finally, it's possible that factors beyond the agent's epistemic state influence their responsibility (**Hypothesis #3:** *"Beyond Epistemic States"*). Some legal theorists and philosophers argue that it not only matters what someone did, but also what they *would have done* had they known about the harmful consequences of their action (Luban, 1998; Wieland, 2017, 2019). Agents are more responsible if they would have acted anyhow (see Nanay, 2010). The underlying 'action model' of what the agent would have done depending on their epistemic state reveals something about their motive to act (Wieland, 2017; Yaffe, 2018). If the agent would have acted anyhow, this suggests 'ill will": a willingness to commit the harm (Wieland, 2017; Yaffe, 2018). While this list of hypotheses is not exclusive – and certainly not exhaustive –, these three hypothesis will mark the scope of the current paper. In the General Discussion, we will address further factors of interest, in particular the role of inferences about an agent's choice not to pursue knowledge (Aboody, Zhou, & Jara-Ettinger, 2021).

## The present study

The intentionality of being ignorant has been argued to impact its moral status. However, it is currently an open question whether an agent who actively refrained from getting to know the consequences of their action is evaluated differently than an agent who just happened to be ignorant. In *Experiment 1*, we test **H1** by investigating people's responsibility judgments and belief judgments to agents who are either willfully or unintentionally, ignorant or knowledgeable about the harmful consequences of their actions. In *Experiment 2*, we investigate how people attribute responsibility when the epistemic uncertainty of these agents is explicitly stipulated (**H2**). Finally, in *Experiment 3* we look at how people evaluate actions from willful and unintentional ignorance if they learn what the agent would have done had they known (**H3**). The experiment pre-registrations, materials, data, and analysis code are available here: `https://github.com/cicl-stanford/father-dont-forgive`.

# Experiment 1: Manipulating ignorance

In this study, we investigate how people hold ignorant versus knowledgeable agents responsible depending on how the they acquired their epistemic state.

## Methods

**Participants & Design** We recruited 201 participants (*age*: M = 37, SD = 13; *gender*: Female = 89, Male = 108, Nonbinary = 3; *race*: Asian = 31, Black/African American = 31, Multiracial = 6, Native Hawaiian/Pacific Islander = 1, White = 145, Other = 6), via Prolific (Palan & Schitter, 2018). The experiment has a 2 'intentionality' (unintentional vs. willful) $\times$ 2 'epistemic state' (ignorance vs. knowledge) design. Participants saw all four epistemic conditions (unintentional ignorance, willful ignorance, unintentional knowledge, willful knowledge). The experiment also included four different scenarios in which these epistemic conditions were presented: 'fertilizer', 'laundry detergent', 'sunscreen', and 'exterior paint'. Each participant was randomly assigned to one of four experimental conditions. Each of the four experimental conditions contained an unique set of four epistemic condition [scenario] combinations, in which each scenario and each epistemic condition appears once (e.g. unintentional knowledge [fertilizer], unintentional ignorance [paint], willful knowledge [laundry], willful ignorance [sunscreen]). The four vignettes were presented in randomized order.

**Procedure** All experiments were programmed with `jsPsych` (de Leeuw, 2015). We instructed participants that this study was about assigning responsibility to agents in various scenarios. They were informed that they would read four short stories about a CEO in a company, answer a couple of questions, and make responsibility judgments. For each scenario, participants first learned that there was a CEO who faces a decision whether to launch a product with potentially negative side effects for the environment (cf. Knobe, 2003). Here is the first part of the 'fertilizer' scenario:

> Taylor is the CEO of a company for farming supplies. The company plans to produce and sell a new fertilizer for agricultural farming, "GreenLine". If launched, "GreenLine" would significantly increase the company's profits. However, this new fertilizer includes a relatively unknown enzyme. It is possible that this enzyme would harm local wildlife if released into the ground. If this were the case, applying "GreenLine" would still effectively fertilize farmland, but also kill small animals like bees, birds, and rodents. The decision is now with Taylor. The new fertilizer will only go into production if Taylor approves it.

The scenario then varies the CEO's epistemic condition. It states that there is a new scientific study showing that the enzyme in "GreenLine" does indeed harm wildlife. The CEO has read a research report that has recently been published summarizing existing research on the new enzyme. We varied whether reading the report allowed the CEO to learn about the

harmful effect, whether the CEO needed to inquire further and, if so, whether the CEO actually did inquire further:

**Unintentional Ignorance** However, the report fails to mention the existence of the new study and what the study finds: the harmful effect of the enzyme in "GreenLine" on wildlife. This means that Taylor does not know that "GreenLine" harms wildlife.

**Willful Ignorance** The report mentions the existence of the new study, but it does not mention what the study finds: the harmful effect of the enzyme in "GreenLine" on wildlife. Taylor decided *not* to look into the new study and learn what the study finds. This means that Taylor does not know that "GreenLine" harms wildlife.

**Unintentional Knowledge** The report mentions the existence of the new study, and it also mentions what the study finds: the harmful effect of the enzyme in "GreenLine" on wildlife. This means that Taylor knows that "GreenLine" harms wildlife.

**Willful Knowledge** The report mentions the existence of the new study, but it does not mention what the study finds: the harmful effect of the enzyme in "GreenLine" on wildlife. Taylor decided to look into the new study and learn what the study finds. This means that Taylor knows that "GreenLine" harms wildlife.

Participants were then asked a set of comprehension check questions about this first part of scenario. These questions asked about the nature of the outcome, the exact content of the research report, the actions of the CEO (reading the report vs. looking further into the study) and the final epistemic state of the CEO. Participants had to answer all questions correctly to proceed. In the second part of the scenario, presented on a separate screen, the company committee comes together to discuss the potential production launch of "GreenLine". In the fertilizer scenario, Taylor approves the product and as a result of the widespread sale of "GreenLine", several wildlife populations are being harmed. The CEO's decision and the outcome of their action was the same in all epistemic conditions.

**Dependent Variables** After reading the two parts of each scenario, participants answered two questions. First, a question about *responsibility*, "To what extent is Taylor responsible for several wildlife populations having been harmed?" and, second, a question about the agent's *belief*, "How likely did Taylor think that several wildlife populations would get harmed?". Participants responded using sliders on 11-point Likert scales with the endpoints labeled "not at all" (0) and "very much" (10).

## Results

Figure 1 shows participants' responsibility and belief judgments as a function of the agent's epistemic state and the intentionality of their epistemic state. Whether the agent knew about the harmful effects or was ignorant about it strongly affected how responsible participants saw the agent (see Table 1): A knowledgeable agent was judged more responsible for the harm than an ignorant one.[1] Likewise, willful

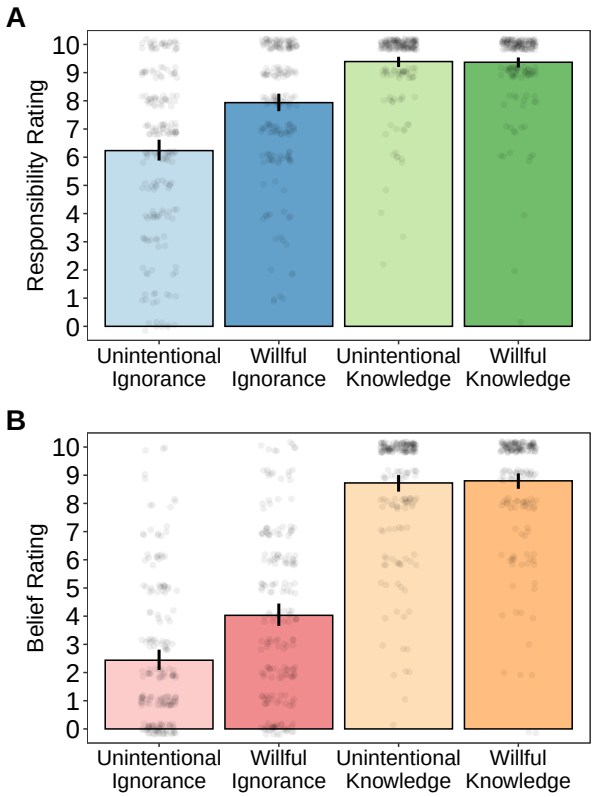


Figure 1: **Experiment 1.** Participants' (a) responsibility and (b) belief judgments as a function of intentionality (unintentional vs. willful) and epistemic state (ignorance vs. knowledge). *Note*: In all of the figures, bar graphs indicate group means, error bars are bootstrapped 95% confidence intervals, and small circles are individual participants' judgments (jittered for visibility).

Table 1: **Experiment 1**. Estimates of the posterior mean and 95% highest density intervals (HDIs) for the different predictors in the Bayesian regression model. Variable coding (used in all experiments): "Epistemic State" $\in \{-1 =$ 'ignorant', $1 =$ 'knowing'$\}$ and "Intentionality" $\in \{-1 =$ 'unintentional', $1 =$ 'willful'$\}$.

| Term | Responsibility | Belief |
|---|---|---|
| Intercept | 8.24 [8.01, 8.48] | 6.00 [5.66, 6.35] |
| Epistemic State | 1.15 [1.03, 1.27] | 2.76 [2.62, 2.91] |
| Intentionality | 0.42 [0.30, 0.54] | 0.42 [0.28, 0.56] |
| Epistemic State:Intentionality | −0.43 [−0.54, −0.31] | −0.38 [−0.53, −0.23] |

[1]We adopt the convention of calling something an effect if the 95% highest density interval (HDI) of the posterior distribution of the estimated parameter in the Bayesian model excludes 0.

agents were judged more responsible than an unintentional ones. There was also an interaction between epistemic state and intentionality. While it did not matter for the knowledgeable agent whether they willfully or unintentionally acquired their knowledge about the outcome, intentionality mattered for ignorant agents. A wilfully ignorant agent was held more responsible than an unintentionally ignorant one.

Participants' responsibility judgments were mirrored by their belief judgments: willfully ignorant agents were judged as more likely to believe that the harm would occur than unintentionally ignorant agents. In contrast, people did not attribute more belief to the willfully versus unintentionally knowledgeable agent.

### Discussion

The results of Experiment 1 show that agents acting out of ignorance are judged less responsible for the negative consequences of their action than an agent who was aware of them. However, in contrast to **H1**, we find that willful and unintentional ignorance are not equally exculpating. People attribute more responsibility to willfully ignorant agents. Unintentionally ignorant agents are, however, not fully exculpated in our scenarios and still receive a substantial degree of blame (Sarin & Cushman, 2023), yet less than unintentionally ignorant agents. In contrast, when agents are knowledgeable, it doesn't matter how they acquired their knowledge. The fact that people also judged a deliberately ignorant agent to believe the harmful consequences of their action to be more likely supports **H2**. People may have attributed more responsibility to the willfully ignorant agent because they inferred a greater degree of belief in the harm from their deliberate ignorance. In Experiment 2, we test whether willfully ignorant agents receive more blame even when their epistemic uncertainty is explicitly stipulated.

## Experiment 2: Manipulating beliefs

This experiment investigates people's responsibility about ignorant and knowledgable agents while explicitly stipulating their degree of belief.

### Methods

**Participants & Design** We recruited 395 participants (*age*: M = 39, SD = 14; *gender*: Female = 179, Male = 206, Nonbinary = 5; *race*: Asian = 30, Black/African American = 31, Multiracial = 21, White = 302, Other = 7), via Prolific. No participants were excluded. The experiment has a 2 epistemic state (knowledge vs. ignorance) $\times$ 2 intentionality (willful vs. unintentional) $\times$ 2 uncertainty (20% vs. 50%) design. Epistemic state and intentionality were manipulated within participants, and uncertainty was manipulated between participants. Otherwise the design was identical to Experiment 1.

**Procedure** The procedure was largely identical to that of Experiment 1, but this time we explicitly stipulated how certain the agent was that the negative outcome is going to occur as a result of their decision to launch the product. In the 'Knowledge' conditions, the CEO thinks the likelihood of the program harming the environment is 100%. In the 'Ignorance' conditions, the CEO thinks the likelihood is either 20% ('low chance') or 50% ('medium chance').

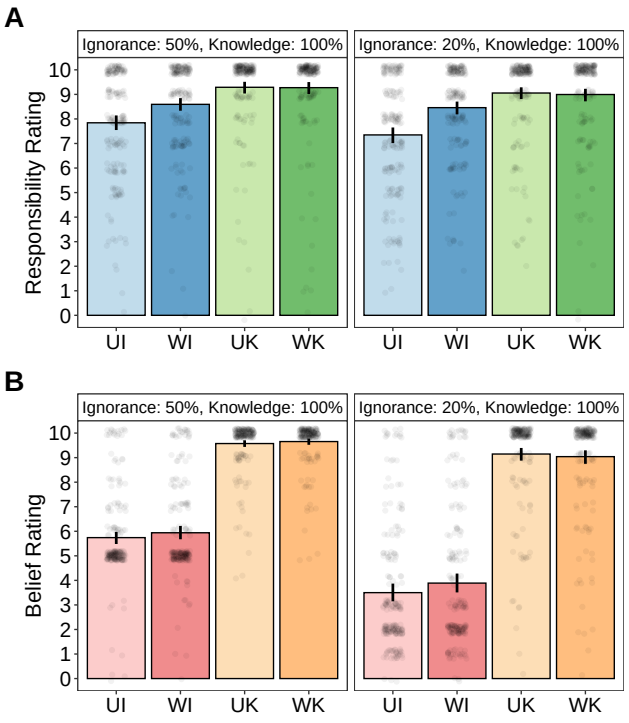


Figure 2: **Experiment 2.** Participants' (a) responsibility judgments and (b) belief judgments as a function of epistemic state (ignorance vs. knowledge), intentionality (unintentional vs. willful), and the ignorant agent's certainty level in the ignorance condition (left: 50%, right: 20%). 'UI' = Unintentional Ignorance, 'WI' = Willful Ignorance, 'UK' = Unintentional Knowledge,'WK' = Willful Knowledge.

**Dependent Variables** In addition to the *responsibility* and *belief* questions, we added the following counterfactual question in the low uncertainty conditions: "How likely is it that [the CEO] would have launched the product had he known that [several wildlife populations would get harmed]?", using the same response scale and labels as for the other questions.

### Results

As Figure 2b shows, we successfully manipulated participants' assumptions about the agent's belief. Unlike in Experiment 1, this time people did not attribute more belief about the outcome to the willfully ignorant compared to the unintentionally ignorant agent (see Table 2). Figure 2a shows that, despite this, willfully ignorant agents were still held more responsible than unintentionally ignorant agents. Willfully and unintentionally knowledgeable agents received the same responsibility. What did, however, make a difference to people's belief attributions to ig-

Table 2: **Experiment 2**. Estimates of the posterior mean and 95% highest density intervals (HDIs) for the different predictors in the Bayesian regression model.

| Term | Responsibility | Belief |
|---|---|---|
| Intercept | 8.61 [8.41, 8.78] | 7.06 [6.89, 7.25] |
| Epistemic State | 0.55 [0.48, 0.61] | 2.30 [2.21, 2.39] |
| Intentionality | 0.22 [0.16, 0.29] | 0.07 [−0.02, 0.16] |
| Epistemic State:Intentionality | −0.24 [−0.31, −0.018] | −0.08 [−0.17, 0.02] |

norant agents was the stipulated level of the agent's uncertainty, $B = 1.07$, 95% HDI $[0.86, 1.28]$. Just considering the 'Ignorance' condition, people attributed a greater degree of belief about the outcome when the agent was said to believe the likelihood of the outcome to be 50% than when it was 20%. However, this difference in belief attributions did not translate into a difference in people's responsibility attributions: While people judged a willfully ignorant agent as more responsible than an unintentionally ignorant one, this was unaffected by whether the agents believed that the chances of the negative outcome were 50% or 20%, $B = 0.15$, 95% HDI $[-0.02, 0.33]$. Finally, we found that the intentionality of ignorant agents impacted participants' counterfactual inferences about what an agent would have done in case of knowledge, $B = 0.28$, 95% HDI $[0.15, 0.40]$. Participants made different inferences from the intentionality of the agent's ignorance: Participants thought that the willfully ignorant agent was more likely to launch the product had they known about the negative effect compared to the unintentionally ignorant agent.

### Discussion

In Experiment 2, we found that people hold a willfully ignorant agent more responsible than an agent whose ignorance was not self-inflicted, even when we explicitly stated the agents' degree of belief. This shows that, contra to **H2**, it's not people's epistemic inferences alone that underpin their increased responsibility attribution to willful versus unintentional ignorance. Interestingly, we also found that there was no quantitative mapping between the degree of uncertainty and people's responsibility judgments. People did not consider an agent who believed the likelihood of the outcome to be 50% more responsible than an agent who only believed it to be 20% (but see Gerstenberg et al., 2018; Johnson & Rips, 2015). All that mattered was whether an agent knew or didn't know. We found tentative evidence for the idea that the inferred action model affected responsibility judgments as predicted by **H3**. People thought that the willfully ignorant would be more likely to launch the product anyways, were they to find out about the harm it would do. To test this hypothesis more directly, we explicitly manipulated the action model in Experiment 3.

## Experiment 3: Manipulating action models

In this study, we investigate people's responsibility judgments to willfully versus unintentionally ignorant agents while manipulating the agents' action models.

**Participants & Design** We recruited 194 participants (*age*: M = 37, SD = 14; *gender*: Female = 106, Male = 86, Non-binary = 2; *race*: Asian = 20, Black/African American = 11, Multiracial = 8, White = 151, Other = 2), via Prolific (Palan & Schitter, 2018). No participants were excluded. The experiment has a 2 action model (dependent vs. independent) $\times$ 2 intentionality of ignorance (unintentional vs. willful) design. Both factors were manipulated within participants.

**Procedure** In this study, participants only got to see the "ignorance" scenarios. For each ignorance condition, we additionally manipulated the causal action model of the agent. To do so, we explicitly stipulated in the first part of the scenario whether the agent would launch the product if they found out about its negative consequences.

> **Dependent** If Charlie were to find out that "SafeSol" does *not* destroy coral reefs, they would decide to launch the production of "SafeSol". If Charlie were to find out that "SafeSol" does destroy coral reefs, they would *not* decide to launch the production of "SafeSol".
>
> **Independent** If Charlie were to find out that "SafeSol" does *not* destroy coral reefs, they would decide to launch the production of "SafeSol". If Charlie were to find out that "SafeSol" does destroy coral reefs, they would **still** decide to launch the production of "SafeSol".

The ignorant agent's epistemic certainty was fixed at 50%. Otherwise, scenarios were identical to those in Experiment 1 and 2.

**Dependent Variables** *Responsibility*, *belief* and *counterfactual* judgments were assessed as in Experiment 1 and 2. Here, we just focus on responsibility and belief questions.

### Results

Figure 3a shows that participants again held willfully ignorant agents more responsible than unintentionally ignorant agents. Participants also judged an agent who would launch the product irrespective of whether they knew about the negative consequences as more responsible than someone who would refrain from launching were they to find out that the product harms the environment (see Table 3). The increase in responsibility attributions to willful ignorance depended on the underlying action model of the agent: When an agent would launch the product even if they were to find out about its negative consequences, the difference in responsibility between unintentional and willful ignorance was much smaller than when the agent's actions depended on knowing about the consequences. We also found small effects of intentionality and action model on people's belief attributions, although people's general attribution of belief was fairly consistent around the mid-point in all four ignorance conditions (see Figure 3b).

### Discussion

In Experiment 3, we find that an agent's action model makes a difference to people's responsibility judgments: Ignorant

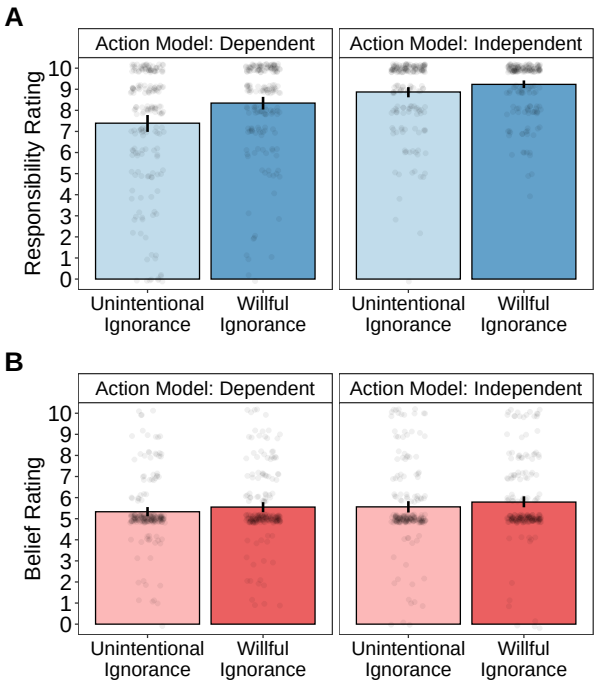


Figure 3: **Experiment 3**. Participants' (a) responsibility judgments and (b) belief judgments as a function of intentionality of ignorance (unintentional vs. willful), and the ignorant agent's action model (dependent vs. independent).

agents who would act irrespective of whether they know or don't know about the harmful consequences of their actions are judged more responsible than agents for whom knowing about the harm would make a difference. While an action model that is independent of the agent's epistemic state reduces the perceived difference in responsibility of a willfully versus unintentionally ignorant agent, it does not make it go away completely: People still judge a willfully ignorant agent as somewhat more responsible than an agent whose ignorance was unintentional.

## General discussion

In three experiments, we explored how people hold agents responsible who differed in what they knew, and in how they had arrived at their knowledge state. We examined three hypotheses about the relationship between knowledge and responsibility: Ignorance is generally exculpating (**H1**), responsibility judgments depend on what an agent believed (**H2**), and responsibility judgments depend on what an agent would have done, had they known about the harm their action would cause (**H3**). For ignorant agents, both inferences about the agents' epistemic states as well as inferences about their action model affected people's responsibility judgments. So while we find some evidence for each hypothesis, even when all of these factors have been accounted for, people still hold willfully ignorant agents somewhat more responsible than unintentionally ignorant agents. In addition, we found that when an agent knew about the negative consequences of their action, they were judged highly responsible and it didn't matter how they had come to know.

What else could make people want to hold willfully ignorant agents accountable for their actions? Some legal frameworks determine an agent's culpability based on whether they remained willfully ignorant with the specific *motive* of supporting their defense in case of prosecution (Hellman, 2009). Accordingly, people may infer that an agent remained willfully ignorant because they wanted to have plausible deniability as an excuse (McGoey, 2012). There might be other reasons, too, for not wanting to know. Maybe a willfully ignorant agent just didn't care enough to find out about the potentially harmful consequences of their action (Sarin & Cushman, 2023)? People might perceive the deliberately ignorant CEO to violate the epistemic duties an agent in their position needs to comply with (Hall & Johnson, 1998). As our study suggests, people make rich inferences from the exact conditions of an agent's ignorance, and these inferences drive significant differences in the moral evaluations of their actions. In general, the choice not to pursue information allows for a variety of inferences about the agent's rewards and values (Aboody, Zhou, & Jara-Ettinger, 2021): they might not value the information enough or simply do not care about the potential harm to make an effort to find out. Moreover, here, we focused on scenarios with negative outcomes. In future work, we'd like to explore what role willful ignorance plays in scenarios with positive outcomes. For example, willful ignorance is evaluated positively, when it serves the function to remain impartial, or to avoid biased decisions (Hertwig & Engel, 2016). Varying the specific reasons an agent might have for avoiding knowledge might allow us to gain further insights into how deliberate ignorance is evaluated.

Table 3: **Experiment 3**. Estimates of the posterior mean and 95% highest density intervals (HDIs) for the different predictors in the Bayesian regression model. Variable coding: "Action Model" $\in \{-1 =$ 'dependent', $1 =$ 'independent'$\}$.

| Term | Responsibility | Belief |
|---|---|---|
| Intercept | 8.45 [8.13, 8.75] | 5.51 [4.95, 5.79] |
| Action Model | 0.59 [0.48, 0.70] | 0.12 [0.01, 0.23] |
| Intentionality | 0.2 [0.22, 0.44] | 0.11 [0.01, 0.21] |
| Action Model:Intentionality | −0.15 [−0.26, −0.03] | 0.00 [−0.09, 0.11] |

## Conclusion

When an agent willfully remains ignorant about potentially negative consequences of their action, people make inferences about why they chose not to know. What inferences people make affects how responsible the agent is viewed. People infer that a willfully ignorant agent was more likely to believe that a negative outcome would happen, and that the agent would have acted anyhow even if they had known. While these inferences partly explain the 'willful ignorance' effect, more work is needed to fully uncover what's going on.

## Acknowledgments

LK, TG, and RZ were supported by a grant from the Binational Science Foundation (BSF) #2020212. LK and TG were supported by a research grant from the Stanford Institute for Human-Centered Artificial Intelligence (HAI).

## References

Aboody, R., Davis, I., Dunham, Y., & Jara-Ettinger, J. (2021). I can tell you know a lot, although i'm not sure what: Modeling broad epistemic inference from minimal action.

Aboody, R., Zhou, C., & Jara-Ettinger, J. (2021). In pursuit of knowledge: Preschoolers expect agents to weigh information gain and information cost when deciding whether to explore. *Child Development*, *92*(5), 1919–1931.

Buchak, L. (2014). Belief, credence, and norms. *Philosophical studies*, *169*(2), 285–311.

Chockler, H., & Halpern, J. Y. (2004). Responsibility and blame: A structural-model approach. *Journal of Artificial Intelligence Research*, *22*, 93–115.

Conrads, J., & Irlenbusch, B. (2013). Strategic ignorance in ultimatum bargaining. *Journal of Economic Behavior & Organization*, *92*, 104–115.

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition*, *108*(2), 353–380.

Dana, J. (2006). Strategic ignorance and ethical behavior in organizations. In *Ethics in groups.* Emerald Group Publishing Limited.

Dana, J., Weber, R. A., & Kuang, J. X. (2007). Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness. *Economic Theory*, *33*(1), 67–80.

de Leeuw, J. R. (2015). jspsych: A javascript library for creating behavioral experiments in a web browser. *Behavior Research Methods*, *47*(1), 1–12.

Edwards, J. L. J. (1954). The criminal degrees of knowledge. *The Modern Law Review*, 294–314.

Gerstenberg, T., & Lagnado, D. A. (2012). When contributions make a difference: Explaining order effects in responsibility attribution. *Psychonomic Bulletin & Review*, *19*(4), 729–736.

Gerstenberg, T., Ullman, T. D., Nagel, J., Kleiman-Weiner, M., Lagnado, D. A., & Tenenbaum, J. B. (2018). Lucky or clever? from expectations to responsibility judgments. *Cognition*, *177*, 122–141.

Gorman, D. E. (2011). Global-tech appliances, inc. v. seb sa: Invoking the doctrine of willful blindness to bring those who lack knowledge of induced infringement within sec. 271 (b)'s prohibition. *Tul. J. Tech. & Intell. Prop.*, *14*, 397.

Hall, R. J., & Johnson, C. R. (1998). The epistemic duty to seek more evidence. *American Philosophical Quarterly*, *35*(2), 129–139.

Hellman, D. (2009). Willfully blind for good reason. *Criminal Law and Philosophy*, *3*(3), 301–316.

Hertwig, R., & Engel, C. (2016). Homo ignorans: Deliberately choosing not to know. *Perspectives on Psychological Science*, *11*(3), 359–372.

Husak, D. N., & Callender, C. A. (2019). Wilful ignorance, knowledge, and the "equal culpability" thesis: A study of the deeper significance of the principle of legality. In *Criminal law* (pp. 203–244). Routledge.

Johnson, S. G., & Rips, L. J. (2015). Do the right thing: The assumption of optimality in lay decision theory and causal judgment. *Cognitive psychology*, *77*, 42–76.

Kirfel, L., & Hannikainen, I. (2022). Why blame the ostrich. *Understanding Culpability for Willful Ignorance [Preprint]. PsyArXiv*, *10*.

Kirfel, L., & Lagnado, D. (2021). Causal judgments about atypical actions are influenced by agents' epistemic states. *Cognition*, *212*, 104721.

Knobe, J. (2003). Intentional action in folk psychology: An experimental investigation. *Philosophical Psychology*, *16*(2), 309–324.

Lagnado, D. A., & Channon, S. (2008). Judgments of cause and blame: The effects of intentionality and foreseeability. *Cognition*, *108*(3), 754–770.

Langenhoff, A. F., Wiegmann, A., Halpern, J. Y., Tenenbaum, J. B., & Gerstenberg, T. (2021). Predicting responsibility judgments from dispositional inferences and causal attributions. *Cognitive Psychology*, *129*, 101412.

Luban, D. (1998). Contrived ignorance. *Geo. LJ*, *87*, 957.

McGoey, L. (2012). The logic of strategic ignorance. *The British journal of sociology*, *63*(3), 533–576.

Nanay, B. (2010). Morality or modality?: What does the attribution of intentionality depend on? *Canadian Journal of Philosophy*, *40*(1), 25–39.

Nicolas, R. (2004). Knowledge management impacts on decision making process. *Journal of knowledge management*.

Nyborg, K. (2011). I don't want to hear about it: Rational ignorance among duty-oriented consumers. *Journal of Economic Behavior & Organization*, *79*(3), 263–274.

Palan, S., & Schitter, C. (2018). Prolific. ac—a subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, *17*, 22–27.

Phillips, D. (2019). *Rossian ethics: Wd ross and contemporary moral theory*. Oxford University Press.

Ross, W. D. (2011). *Foundations of ethics*. Read Books Ltd.

Sarch, A. (2018). Willful ignorance in law and morality. *Philosophy Compass*, *13*(5), e12490.

Sargent, R. H., & Newman, L. S. (2021). Pluralistic ignorance research in psychology: A scoping review of topic and method variation and directions for future research. *Review of General Psychology*, *25*(2), 163–184.

Sarin, A., & Cushman, F. A. (2023, Jan). *Punishment in*

*negligence is multifactorial: influenced by outcome, lack of due care, and the mere failure of thought.* PsyArXiv.

Shepperd, J. A., & Howell, J. L. (2015). Responding to psychological threats with deliberate ignorance: Causes and remedies. In *Handbook of personal security* (pp. 275–292). Psychology Press.

Smith, H. (1983). Culpable ignorance. *The Philosophical Review*, *92*(4), 543–571.

Wieland, J. W. (2017). Responsibility for strategic ignorance. *Synthese*, *194*(11), 4477–4497.

Wieland, J. W. (2019). Willful ignorance and bad motives. *Erkenntnis*, *84*(6), 1409–1428.

Yaffe, G. (2018). The point of mens rea: The case of willful ignorance. *Criminal Law and Philosophy*, *12*(1), 19–44.

Young, L., & Tsoi, L. (2013). When mental states matter, when they don't, and what that means for morality. *Social and personality psychology compass*, *7*(8), 585–604.