Mens rea ascription, expertise and outcome effects: professional judges surveyed

Markus Kneer, Sacha Bourgeois-Gironde

MENS REA ASCRIPTION, EXPERTISE AND OUTCOME EFFECTS: PROFESSIONAL JUDGES SURVEYED

Markus Kneer
University of Pittsburgh

Sacha Bourgeois-Gironde
Institut Jean Nicod, Paris

Abstract

A coherent practice of *mens rea* ascription in criminal law presupposes a concept of *mens rea* which is insensitive to the moral valence of an action's outcome. For instance, an assessment of whether an agent harmed another person *intentionally* should be unaffected by the *severity* of harm done. Ascriptions of intentionality made by laypeople, however, are subject to a strong outcome bias. As demonstrated by the Knobe effect, a knowingly incurred negative side effect is standardly judged intentional, whereas a positive side effect is not. We report the first empirical investigation into intentionality ascriptions made by professional judges, which finds (i) that professionals are sensitive to the moral valence of outcome, and (ii) that the worse the outcome, the higher the propensity to ascribe intentionality. Our results thus suggest that the intentionality ascriptions of professional judges are inconsistent with the concept of *mens rea* supposedly at the foundation of criminal law.

1. Introduction: The Knobe Effect and Criminal Jurisprudence

1.1 Two Concepts of Intentionality

Consider Knobe's well-known CHAIRMAN scenario: The chairman of a company is approached by his advisor, who recommends a new business strategy. The strategy is expected to increase profits and to harm the environment. The chairman responds that he does not care about the environment and gives his advisor the green light. Everything turns out as predicted: Profits increase and the environment suffers. Did

the chairman harm the environment intentionally? The overwhelming majority of philosophically uninitiated people judge the foreseen negative side effect intentional. But faced with identical cases that differ only in so far as the outcome is not negative but positive (i.e. the environment benefits from the new strategy), the side effect is predominantly judged as a nonintentional by-product of the main action. The asymmetry – frequently called the 'Knobe effect' – has been widely replicated (Knobe 2003a, 2003b, 2004; Mele and Cushman 2007; for survey articles, cf. Feltz 2007b; Cova 2016). The effect is found robustly across different cultures (Knobe and Burra 2006; Dalbauer and Hergovich 2013) and ages (Leslie et al. 2006). It extends to a wide range of ascriptions of mental states such as desire (Tannenbaum et al. 2007), knowledge (Beebe and Buckwalter 2010; Beebe and Jensen 2012), belief (Beebe 2013; Kneer forthcoming) and attributions of non-mental properties such as causal involvement (Knobe and Fraser 2008).

The folk concept of intentionality, this suggests, is sensitive to moral valence – it is morally, or normatively, *charged*. If the outcome is negative, foreknowledge standardly suffices for people to ascribe intentionality, if it is positive, foreknowledge does standardly not suffice. ('Standardly' since the introduction of further factors such as agent regret can disrupt the asymmetry, cf. Phelan and Sarkissian (2008), Cushman and Mele (2008) and Cova et al. (2012)). The folk concept differs from what we will call the *clinical* concept of intentionality, i.e. the concept prevalent in law and philosophy. On this view, intentionality involves both a cognitive element, i.e. awareness or knowledge of the consequences, and a conative element, i.e. a desire or other pro-attitude to bring about the envisioned consequences. (Butler 1978; Katz 1987; Mele 1992; Moore 2011; Adams 2015). For an action to count as intentional, both elements are necessary,

independently of the moral valence of the outcome (for dissenting views cf. Harman (1976) and Lowe (1978), for comparative discussion across law and philosophy, cf. Duff (1989)). Criminal law standardly invokes the clinical concept of intentionality. The US Model Penal Code (section 2.02), for instance, distinguishes explicitly between the *mens reas* intentionality (or purpose) and knowledge (the agent's awareness that his actions will produce a certain result). But this distinction could not be upheld in an unqualified fashion if knowledge was sometimes sufficient for intentionality, as the Knobe effect suggests.

1.2 The Mismatch between Folk Psychology and the Law

The foregoing discussion suggests a severe mismatch between the concept of intentionality at the foundation of criminal law on the one hand, and the folk concept of intentionality on the other. Citizens might thus misinterpret the law, question the verdicts of high-profile trials and challenge the law's legitimacy more generally (Tyler 2006; P. H. Robinson and Darley 1995). In Anglophone jurisprudence, where laypeople juries attribute *mens rea*, the mismatch is particularly problematic: While the law draws a clear, outcome-independent distinction between the *mens reas* of intentionality and knowledge, on the folk view knowledge can suffice for the ascription of intentionality. If this is the case only, or predominantly, with respect to side effects, then those taken to trial for harmful side effects are judged by different standards than those charged for harmful main-effects. In the latter case, foreknowledge is *not* sufficient for intentionality ascription (and thus the most severe punishment), in the former case, it is. Note also that a small, though significant minority of laypeople employ a clinical concept of intentionality. This, too, challenges the principle of a fair and equal trial for all: Defendants who have acted with mere foreknowledge (i.e. without a pro-attitude towards the side-effect) will be attributed the *mens rea* of knowledge by juries holding the clinical

view, others will be attributed the more inculpating *mens rea* of intentionality by juries employing the normatively charged concept of intentionality. Advocates of a strict distinction between intentionality and foreknowledge are thus concerned that defendants who act with mere foreknowledge might frequently be judged and punished too harshly.[1]

Perhaps, one might think, the impact of the mismatch just described is exaggerated: The fact that central legal and folk concepts differ does not mean that the folk cannot grasp, or – under careful instruction as is common practice in criminal trials – ascribe *mens reas* as defined by the law. There is a small empirical literature that investigates whether the legally uninitiated can competently distinguish the *mens reas* laid out in the US model penal code, and whether they can rank them appropriately in terms of culpability and punishment. Experiments by P. H. Robinson and Darley (1995) suggest that, by and large, they can. The majority of studies (Severance et al. 1992; Levinson 2005; Shen et al. 2011; Ginther et al. 2014) however, report that the folk have considerable difficulties in reliably distinguishing the different *mens rea* concepts and in ranking their respective culpability in ways consistent with the Model Penal Code. What is more, the provision of jury instructions are standardly found to be of little help, which might be

[1] A worry: What drives the Knobe effect are differently valenced outcomes. But – one might argue – the distinction between positive and negative moral or normative valence is mute as regards legal matters, since the only outcomes of relevance are negative ones. One doesn't get taken to court for exemplary behaviour, but for breaking the law, standardly associated with doing harm or damage. Though there might thus be an asymmetry across positive and negative outcomes, the fact that only the latter matter ensures equality before the law: Those doing harm do not get judged differently from those doing good, because the latter don't get judged in court in the first place. As the main discussion should make clear, however, this worry misses the mark. The problematic here addressed arises *not* from the asymmetry of intentionality judgments across differently valenced *outcomes*, but from potentially different *concepts* of intentionality at work in criminal law – one that requires a conative attitude besides foreknowledge, and another one which does not. Differently put, the problem arises from the fact that a clear distinction between the *mens reas* of intentionality and knowledge is not guaranteed in similarly, that is, negatively valenced, cases.

one of the reasons why jurors so frequently ask for clarifications of *mens rea* concepts in criminal trials (Lacey 1993).

Let's briefly take stock: The Knobe effect reveals a serious mismatch between the normatively charged folk concept of intentionality and the clinical concept of intentionality prevalent in criminal law. The mismatch matters both theoretically and practically, since the legally uninitiated have difficulties adapting to the clinical concept in contexts of criminal jurisprudence. In order to better understand the conceptual conflict, and devise ways to address it, the next sections explore the Knobe effect and its implications for the nature of intentional action in more depth.

1.3 Competence v. Bias Accounts

The Knobe effect has sparked extensive debate as to whether the normatively charged concept captures the nature of intentionality better than the clinical one that dominates the philosophical literature and the law (for reviews see Feltz (2007a); Pettit and Knobe (2009); Cova (2016)). Certain scholars argue that the Knobe effect constitutes a *bias,* and that the folk use of intentionality is frequently distorted (Adams and Steadman 2004; Nadelhoffer 2004a, 2004b, 2006; Alicke 2008; Alicke and Rose 2010; Sauer and Bates 2013). In contrast to such views, several scholars have argued that the Knobe effect testifies to people's *competence* in intentionality ascriptions (cf. e.g. Machery 2008; Hindriks 2008; Pettit and Knobe 2009; Knobe 2010b; Uttich and Lombrozo 2010). According to Knobe, for instance, intentionality ascriptions are sensitive to moral concerns since the *concept* of intentionality itself is constitutively tied to moral features. According to Uttich and Lombrozo (2010), the conscious violation of salient norms such as protecting the environment constitutes *evidence* in favour of certain mental states such as intentionality, whereas norm-

conformance does not. The view differs from Knobe's in so far as it invokes a clinical *concept* of intentionality, whose *application* is deemed sensitive to moral and conventional norms. It differs from bias accounts, since the evaluation of behaviour *vis-à-vis* salient norms is considered an epistemically rewarding, and hence rational, feature of mindreading.

Advocates of competence accounts are inclined to find fault with the law and propose a revision of the legal concept of intentionality (Kobick 2010; Duff 2015). Suggestions of this sort echo an influential article by Malle and Nelson (2003), who argue that when central legal and folk concepts are at odds, the law should adopt the latter so as to foster 'clarity of *mens rea* concepts and a reconciliation of the legal and the layperson's view of human behaviour' (2003: 563). It bears emphasis, however, that this strategy is only sensible if the folk concepts of *mens rea* are sufficiently *uniform* and *systematic*, so as to allow a coherent and reliable practice of *mens rea* attribution. Drawing on Malle and Knobe (1997, 2001), Malle & Nelson argue that most people do indeed converge on a single concept of intentionality (uniformity is thus satisfied), and that said concept 'is systematic in that the judgments are predictable from five core components – belief, desire, intention, awareness, and skill.' (2003: 574).

The proposal of adopting folk concepts of *mens rea* for legal purposes can be challenged on two grounds: First, even if sufficiently uniform and systematic, the lay notion of intentionality might still be considered philosophically confused and thus unfit for legal purposes (Adams 2015). Second, one might have doubts about the uniformity and systematicity of the folk concepts of *mens rea*. Uniformity is under pressure since a significant minority does *not* manifest a side-effect effect with respect to intentionality. This suggests that there are *multiple* folk concepts of intentionality – Cushman & Mele (2008), for instance,

identify 'two and a half' such concepts, Lanteri (2012) counts even more. Similar worries regarding uniformity arise for the ascription of the *mens rea* of knowledge, where a significant minority is not susceptible to the epistemic side-effect effect, cf. Beebe and Buckwalter (2010) as well as Beebe and Jensen (2012).

As advocates of bias accounts are quick to point out, the Knobe effect casts doubt on the systematicity of the folk concept of intentionality: When negative side-effects are at stake, desire – one of Malle & Nelson's core components of intentionality – does not seem to play a role, whereas when positive side-effects or main effects are under consideration, it does. What is more, evidence by Nadelhoffer (2006) demonstrates that moral factors independent of outcome valence such as the character of the defendant and victim – certainly not among the core components of the concept of intentionality – can affect the ascription of intentionality and foresight. Nadelhoffer thus suggests to subsume the Knobe effect under Alicke's culpable control model, whose explanatory scope extends beyond side-effects. We consider this a plausible move (for discussion see Alicke (2008); Nichols and Ulatowski (2007); Cole Wright and Bengson (2009); Knobe (2010b)), and will briefly outline the culpable control model, as it can serve as a theoretical framework for our experiments.

Alicke (1992, 2000, 2008) questions standard moral and legal theories of blame, according to which a *ceteris paribus* increase in personal control warrants an increase in blame (schematically: control→blame). Instead, he argues, the desire to blame an agent sometimes incites blame 'amateurs' (2008:180) to exaggerate evidence regarding the agent's personal control (schematically: blame→control). Blame-validation operates on three types of personal control corresponding to the structural links holding between mental states, behavior and consequences which characterize an action: *Volitional behavior*

control (the mind-behavior link), *causal control* (the behavior consequence link) and *volitional outcome* control (the mind-consequence link), which – roughly – correspond to intentionality, causation, and foresight (Alicke and Rose 2010). The Knobe effect can be interpreted as evidence for exaggerated ascriptions of volitional behavior control, the epistemic side-effect effect and Nadelhoffer's (2006) results about foresight as instances of biased assessments of volitional outcome control. There is also ample evidence consistent with the hypothesis that blame validation increases causal control ascriptions (Alicke 2000; Knobe and Fraser 2008; Hitchcock and Knobe 2009; Cushman 2010).

1.4 Measures to address the Mismatch

All parties to the debate agree that the mismatch between legal and folk concepts of intentionality and other *mens reas* constitutes a problem, in particular in countries with a juror system. Competence theorists argue that the law must adopt the folk concepts. Bias theorists, by contrast, might want to advocate the abolishment of juries composed of laypeople: Legal professionals who are well-versed with the law and its requirements, and who have received extensive training, one might suppose, are less susceptible to outcome biases such as the Knobe effect. A proposal of this sort could parallel the *expertise argument* which analytic 'armchair' philosophers have employed to call the philosophical import of experimental studies about folk intuitions into question. Just as the intuitions and judgments of mathematicians, expert chess players or physicists, in their areas of competence, are more reliable and less susceptible to bias than those of laypeople, so are the intuitions of analytic philosophers when it comes to conceptual analysis. (In philosophy, advocates of this view include Kamm (1993); Kauppinen (2007); Ludwig (2007); Williamson (2008, 2011); for critical discussion and empirical studies regarding the

expertise argument, cf. for instance Weinberg et al. (2010); Schulz et al. (2011); Machery (2012); Schwitzgebel and Cushman (2012, 2015); (Alexander 2016)). Analogously, one might argue, the clinical concept of intentionality, dominant in legal scholarship and firmly entrenched in criminal law all over the world, is the appropriate one, since it is the product of expert judgment and rigorous conceptual analysis. What is more, its application, in the hands of experts – that is, professional judges rather than jurors or 'blame amateurs' as Alicke calls them – will be insensitive to distortive factors such as moral or normative valence. Testing this empirical hypothesis constitutes the central goal of this article: We present the first experimental studies about whether the intentionality ascriptions of professional French judges manifest an outcome effect (in France laymen juries are extremely rare).

The results will be of interest independently of whether one favours a bias or a competence account of the Knobe effect and similar phenomena. If the concept of intentionality employed by professionals turns out to be the clinical one, bias theorists could deploy an expertise argument in favour of institutional change. If, on the other hand, experts, too, operate with a charged concept of intentionality, competence theorists might take this as evidence against bias accounts and argue that it is high time to bring the letter of the law into accord with the concepts of laypeople *and* experts. Before reporting the experiments, we'll address a few preliminary worries about what has been said so far.

1.5 Potential Worries

The above considerations might be called into question on two grounds. *First*, the asymmetry in intentionality ascriptions regards side effects but, one might suppose, these do not play an important role in jurisprudence. This is incorrect, since many legal cases focus explicitly

on side effects.[2] What is more, though side-effect scenarios facilitate the experimental investigation, the phenomenon is not peculiarly limited to such cases only – it arises, for instance, also with regards to differently valenced *means* (Cova and Naar 2012), and the very *classification* of goals, means and side effects itself (Ulatowski 2008, 2012; Knobe 2010a).

Second, various accounts of the Knobe effect challenge the dominant view according to which intentionality ascriptions are affected directly by moral or normative considerations. Instead, critics of this view argue, differently valenced side-effects engender asymmetric attributions of outcome-related *desires* (Guglielmo and Malle 2010), *beliefs* (Alfano et al. 2012), perceived *norm-violations* (Uttich and Lombrozo 2010; Lombrozo and Uttich 2010; Alfano et al. 2012) deeply held *values and principles* (Sripada 2010, 2012; Sripada and Konrath 2011), or *attention* paid to the possible consequences (Scaife and Webber 2013). The difference in ascriptions of this sort is, in turn, taken to explain the asymmetric ascription of intentionality.

These are interesting models, and we see no need to pick and choose, as their adequacy has no bearing on our central argument. We are exploring the question whether moral or normative considerations have an impact on intentionality ascriptions in jurisprudential contexts – no matter whether these are *mediated* by ascriptions of conative states, epistemic states, deeply held values or other factors. Furthermore, whereas some (though not all) of the factors mentioned above have indeed proven statistically significant mediators of outcome valence on intentionality ascription, not a single one has been established as

[2] For two recent US Supreme Court cases see *Shell Oil Co. v. United States* (2009) and *Babbitt v. Sweet Home Chapter of Communities for a Great Oregon* (1995), discussed in Kobick and Knobe (2009). Cf. also the plethora of cases regarding discriminative intent in Kobick (2010), where side-effects nearly always play a central role.

the *only* significant mediator. Importantly, even when controlling for the above attitudes in mediation analyses, the *direct* impact of normative considerations remains significant (Cova et al. forthcoming). This leads Cova et al. to conclude that 'moral evaluations still play an irreducible role in shaping our judgments of intentionality' (forthcoming: 12). The focus of this article does not consist in adjudicating between these different views, but in investigating whether the judgments of legal professionals also manifest the Knobe effect. Since blame plays a central role in several explanations of the Knobe effect (e.g. Adams and Steadman (2004); Nadelhoffer (2004a, 2006); for discussion cf. Knobe (2006); Sauer and Bates (2013)) as well as more general accounts such as Alicke's (2000, 2008) culpable control model, we have also measured blame ascriptions in our experiments. To the experiments we will now turn.

2. First Experiment: Good v. Bad Outcomes

The first experiment investigates whether professional judges employ a clinical or a morally charged concept of intentionality.

2.1 Method

2.1.1 Participants

36 professional French judges (23 of whom were female) completed an unpaid online questionnaire. About three times as many were contacted directly via email. Sample size was determined by the number of judges who responded to the invitation and filled out the questionnaire. All complete data sets were used. 26 judges served at a *jurisdiction de première instance* (the lowest type of court), 9 at a *cour d'appel* (court of appeal) and 1 as *conseiller juridique* (legal advisor). 24 participants listed criminal law as their speciality, 17 civil law, 5 social law, and 1 administrative law and 7 other specialities (multiple answers possible). The average professional experience was just under 17 years,

ranging from less than a year to 37 years. The native language of all participants was French, nobody stated familiarity with the Knobe effect and only one participant stated familiarity with experimental philosophy.

2.1.2 Materials and Procedure

In a within-subjects design run with Qualtrics online software, participants were presented with both conditions of Knobe's CHAIRMAN scenario. The order of presentation was randomized. Participants were asked a forced-choice yes/no question whether the chairman had harmed/helped the environment intentionally. In a follow-up question they had to report their dis/agreement with the statement 'The chairman harmed/helped the environment intentionally' on a 7 point Likert scale ranging again from (1) 'strongly disagree' to (7) 'strongly agree'. Finally, participants were asked whether they thought the chairman deserves praise, blame or neither for his action (cf. Appendix 1 for vignettes and questions in French).

2.2 Results

The forced-choice responses differed significantly across conditions (harm v. help) for both orders of presentation. Overall, 86% ascribed intentionality in the harm case, whereas only 11% of the participants did in the help case (McNemar exact test for N=36, $p<.001$). 78% of the respondents gave distinct – that is, *inconsistent* – answers across the two cases. The order of presentation was insignificant, Pearson $\chi^2(1)=.122$, $p=.727$ for *harm* and $\chi^2(1)=.892$, $p=.345$ for *help*. 84% of the participants receiving harm first ascribed intentionality in the harm scenario whereas only 16% did in the help scenario (McNemar exact test for N=19, $p<.001$). Amongst participants receiving the help case first, 88% ascribed intentionality in the harm scenario and 6% in the help scenario (McNemar exact test for N=17 $p<.001$). Counting first choices

only (N=36) to mimic the between-subjects design customary in previous side-effect studies, 84% of the participants attributed intentionality in the harm condition, and 6% in the help condition (Pearson $\chi^2(1)=22.087$, $p<.001$). Figure 1 represents graphically that the Knobe effect is at least as pronounced for professionals (first choices only) as for laypeople (N=78) tested by Knobe (2003a). For a replication with French laypeople cf. Cova & Naar (2012, experiment 1), whose findings are very similar to Knobe's.

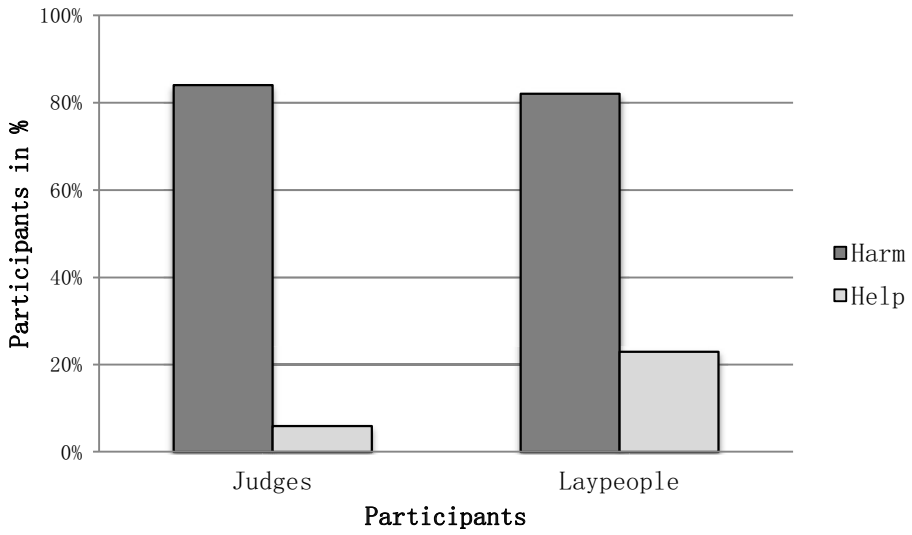

Figure 1: Participants agreeing with the claim that the chairman intentionally harmed/helped the environment for different subject pools. Laypeople data from Knobe (2003a).

On a 7-point Likert scale, mean agreement with the claim that the chairman *harmed* the environment intentionally was 5.67 (SD=1.45), mean agreement with the claim that the chairman *helped* the environment was 2.56 (SD=1.48), cf. Figure 2. A paired-samples t-test reveals the difference to be strongly significant, $t(35)= 8.383$, $p<.001$, 95% CI [2.36;3.87], Cohen's $d=1.40$, a large effect (Cohen 1988). Counting first responses only, thus mimicking a between-subjects design customary for side-effect studies, increases the asymmetry somewhat: The mean agreement for harm is 5.74 (SD=1.33), for help it is

2.18 (SD=1.47). An independent samples t-test revealed a significant difference: t(34)=-7.645, p<.001, 95% CI [-4.51;-2.61], Cohen's *d*=2.55, a very large effect. The findings for professional French judges, both as regards forced-choice and Likert-scale responses replicate those for laypeople. Given that the Knobe effect is similarly pronounced in both subject pools, we can conjecture that both professional judges and laypeople employ a morally charged concept of intentionality. The percentage of those willing to ascribe blame in the harm scenario (blame: 83%, no blame: 0%, neither: 17%) considerably exceeds the one for the help scenario (blame: 20%, no blame: 6%, neither: 74%). The difference for first responses (N=36) is significant, Pearson $\chi^2(2)=11.26$, p=.004).

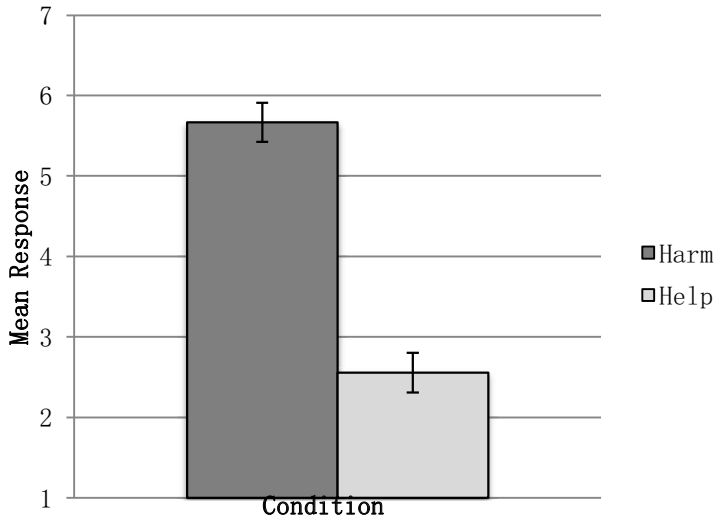

Figure 2: Average intentionality ascriptions in the CHAIRMAN scenario, error bars designate standard error of the mean

2.3 Discussion

The results are loud and clear: The Knobe effect is just as pronounced for professional judges as for laypeople, which suggests that both

groups operate with a morally charged concept of intentionality.[3] The reported asymmetry in blame ascriptions is consistent with models according to which blame plays a central role in intentionality ascriptions, or the ascription of culpable control more generally. If matters were left here, the implications both for social psychology and for the law would be unclear. Competence theorists might argue that the fact that experts, too, employ a charged concept of intentionality implies that lay judgments are innocuous and that the legal definition of intentionality must be brought into accord with the concept used by both laypeople and experts. Bias theorists, by contrast, might argue that – just as in other domains – expert judgments are not always immune from bias (cf. e.g. Schwitzgebel and Cushman (2012, 2015)). What the results demonstrate, on this view, is that the intuitions of legal professionals are *also* distorted. Consequently, the worrisome conceptual mismatch between the letter of the law and the practice of *mens rea* ascription afflicts legal systems where trials are decided by professionals just as much as juror systems.

Which of the two views one is inclined to take depends on a variety of background assumptions regarding *inter alia* the nature of bias (for a review, see Hahn and Harris 2014), one's preferred account of the Knobe effect, and the systematicity and uniformity of folk-psychological concepts. Some of these debates, we would like to suggest, can be circumvented by addressing the topic from a novel angle. Suppose the Knobe effect standardly conceived captures just two data points of a broader phenomenon. Past empirical research on intentionality ascriptions might have inappropriately focused on clear

[3] In an experiment with 59 professional French lawyers, Kneer and Bourgeois-Gironde (forthcoming) report similar results. Participants received both the harm and the help vignette of Knobe's CHAIRMAN scenario. 59% ascribed intentionality in the harm case, while only 21% ascribed intentionality in the help case (McNemar test, $p<.001$). Importantly, despite receiving both conditions (the order was randomized), 59% of all participants judged the two cases differently.

opposites, i.e. pairs of cases contrasting negative with positive outcomes. Rather than testing for such opposites (good v. bad) we might run experiments with *graded* negative outcomes (somewhat bad v. very bad). It could turn out that the willingness to ascribe intentionality is positively correlated with the *severity* of a negative outcome. If this were the case, the competence accounts and the recommended legal revisionism would lose much of their plausibility: It is one thing to advocate a view according to which foresight suffices for the ascription of intentionality if the consequence is harmful. It is quite another to advocate a view according to which the propensity of *mens rea* ascriptions should be commensurate with severity of outcome. On the first view, whether or not I broke a vase intentionally depends, *inter alia*, on whether I foresaw the consequence and on *whether* the consequence was undesirable. On the second view, it depends, *inter alia*, on whether I foresaw the consequence and on *how severe the damage was*. That is, whether or not my action was intentional depends, *inter alia,* on the *value of the vase*. Note that if there were a severity effect of the kind envisioned here, then the Knobe effect could be understood as just a special case of this broader phenomenon. Such an interpretation might cast doubt on a considerable number of explanations of the Knobe effect, which conceive of it in bivalent terms. More importantly for our purposes, if the *mens rea* judgments of laypeople and professionals were susceptible to a severity effect, it is hard to see how the latter could *not* be conceived as a bias, or how the law could reasonably be amended to make sense of it in the first place. For each crime, the law would have to specify minimally required levels of harm that warrant the ascription of guilty states of mind.

Kneer (in preparation) reports data according to which folk judgments concerning the *mens reas* of intentionality, knowledge and

recklessness manifest a severity effect. Our next experiment addresses the question whether the judgments of legal experts are also sensitive to severity of outcome.

3. Experiment 2: Somewhat bad v. very bad outcomes

In comparison to Experiment 1 which invoked clear opposites (good v. bad outcomes), Experiment 2 tested whether the judgments of professional French judges are sensitive to the severity of bad outcomes.

3.1 Method

3.1.1 Participants

32 professional French judges (17 female) completed an unpaid online questionnaire. About three times as many judges were contacted directly by email. Sample size was determined by the number of judges who responded to the invitation and filled out the questionnaire. All complete data sets were used. None of the participants who had been invited to participate in the first study were re-contacted. 24 judges served at a *juridiction de première instance* (the lowest type of court), 5 at a *cour d'appel* (court of appeal) and 3 at the *Cour de Cassation* (the court of cassation, the highest court in France). The average professional experience was around 16 years, ranging from less than a year to 38 years. 20 participants listed civil law as their speciality, 19 criminal law, 4 social law, 2 commercial law, 1 administrative law and 2 other specialities (multiple answers possible). The native language of all participants was French, nobody stated familiarity with the Knobe effect, one participant stated familiarity with experimental philosophy.

3.1.2 Materials and Procedure

In a between-subjects design, participants were randomly assigned one of two conditions of the BEACH TOWN scenario (cf. Appendix 2 for the French version), in which the mayor's actions trigger either a somewhat bad side-effect, or a very bad one:

> The mayor of a small beach town is approached by his advisor who says: "We could build a new highway connection. This would make car traffic much more efficient. However, there would be [minor/severe] adverse effects on the environment. During construction, the animals in the construction zone will [be disturbed/die]. This is [only temporary/not a temporary condition], [everything goes/things will not go] back to normal once construction is finished."
>
> The mayor responds: "I don't care at all about the environment. All I care about is making car traffic as efficient as possible. Let's build the new highway connection."
>
> They build the new highway connection. The animals in the zone are [temporarily disturbed/die]. [Everything goes/Things do not go] back to normal after construction is finished.

Participants were asked to what extent they agreed that the mayor intentionally harmed the environment. Responses were collected on a 7-point Likert scale ranging from 1 (completely disagree) to 7 (completely agree). On a separate screen, participants were asked to what extent they deem the mayor blameworthy for his action (1= not at all, 7= very much).

3.2 Results

The average intentionality ascription level for the *really bad* condition (M=5.18, SD=1.74) exceeds that for the *somewhat bad* condition (M=3.33, SD=1.76), cf. Figure 3. The difference is significant, t(30)=-2.97, p=.006, 95% CI [-3.11; -0.58], Cohen's d=1.06, a large effect size . The values are near-identical for blame ascriptions (really bad: M=5.59, SD=1.18; somewhat bad: M=3.53, SD=1.73), the difference is again significant; t(30)=-3.98,p<.001, 95% CI [-3.11; -1.00], Cohen's d=1.40, a large effect size.

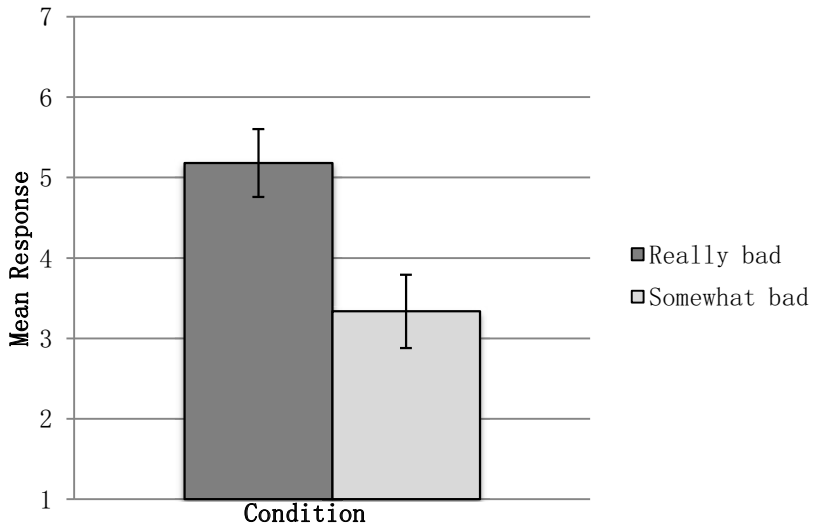

Figure 3: Average intentionality ascriptions for the BEACH TOWN scenario across negative side-effect conditions differing in severity of outcome. Error bars designate standard error of the mean.

3.3 Discussion

For professional French judges, severity of outcome correlates positively with the propensity to ascribe intentionality. This is an important finding in the debate about the impact of action outcome on intentionality ascriptions. It lends support to the hypothesis that the Knobe effect might be but a special case of a broader severity effect. The asymmetric blame ascriptions across the two conditions is consistent with Alicke's culpable control model. An increased attribution of

intentionality (i.e. volitional behaviour control) might be driven by an increased desire to blame the agent. The results are of particular significance for criminal jurisprudence: Contrary to standard legal doctrine, the strict differentiation between the *mens reas* of knowledge and intentionality disintegrates with increasing severity of outcome, which puts pressure on the principle of uniform and equal trial conditions for all.

4. General Discussion

Two experiments with independent groups of professional French judges revealed that their ascriptions of intentionality – the *mens rea* associated with the highest culpability – are susceptible to the Knobe effect and the newly introduced severity effect. More particularly, an asymmetry in intentionality attributions arises both for scenarios differing in outcome valence (Experiment 1, contrasting good v. bad outcomes) and across negative outcomes of different severity (Experiment 2, contrasting somewhat bad with very bad outcomes).

Those who consider the severity effect a bias might worry that defendants who bring about more severe side effects will be punished too harshly, as the less inculpating *mens rea* of knowledge effectively drops out of the picture. Those who advocate a competence account, by contrast, face a new challenge. While such views are not implausible with respect to the Knobe effect strictly conceived, the severity effect requires considerable revision. Take the increasingly popular norm-based views (Uttich and Lombrozo 2010; Holton 2010; Alfano et al. 2012; B. Robinson et al. 2015) as an example, which (roughly) explain the asymmetry in intentionality ascriptions in the CHAIRMAN studies as due to a norm violation in the harm condition, and norm-conformance in the help condition. However, in the BEACH TOWN experiment, both conditions invoke a violation of a single, identical

norm (roughly: do not harm the environment). The pronounced asymmetry in intentionality ascriptions (below the Likert scale midpoint in the moderate condition and above the midpoint in the severe condition) seems to be due to severity of outcome, not a bivalent factor such as norm-conformity v. norm-violation.

Naturally, competence accounts of the Knobe effect could be extended to accommodate severity of outcome on a theoretical level. But the main practical problem, the gap between the letter of the law and the diverging actual practice of *mens rea* ascription, whether effected by laypeople or professional judges, might persist. Though it is relatively straightforward to change the law so as to accommodate the Knobe effect (for suggestions, cf. Kobick (2010) and Duff (2015)), it is less evident how the law could adopt a severity-sensitive concept of intentionality without generating large-scale inner-systematic incoherence. Take, for instance, Coke's Law, a principle at the foundation of nearly every system of criminal law in the world, according to which the assessment of *mens rea* and *actus reus* (the 'guilty act' or material element) should be strictly independent from one another. Such independence in evaluation of *mens rea* would be impossible if it were legally correct to look to the severity of outcome in order to establish whether foreknowledge does or doesn't suffice for the ascription of intentionality.[4]

Future inquiry into *mens rea* ascriptions should pursue a variety of issues that could not be addressed in this article. One important question regards whether the severity effect proves robust across different

[4] To prevent confusion, let it be clear that the measure of *punishment* obviously can take into account the severity of outcome. But the question of whether the defendant in fact had a guilty state of mind is conceptually independent and procedurally prior to the determination of punishment. Except in cases of strict liability, culpability and punishment can only be assessed once both *actus reus* and *mens rea* have been established.

expressions relating to intentionality. While advocates of the 'simple view' postulate a tight conceptual connection between the adjective 'intentionally', the verb 'intend' and the noun 'intention' (e.g. Adams, 1986, 2015; McCann, 1986, 1989, 1991), scholars such as Duff (1986, 2015) and Kobick (2010) question it. Second, it should be explored how professionals respond when explicitly prompted to 'judge as if in court', since they might operate with multiple concepts of intentionality, depending on context. Third, though our findings speak to the experimentally revealed concept of intentionality of professional judges, it would also be helpful to assess what concept of intentionality they explicitly endorse when directly asked about it. Consistency between the experimentally revealed and explicitly endorsed concepts would support a competence view; inconsistency would speak in favour of a bias view since the judges would consider the results here presented as afflicted by error *vis-à-vis* their *own* concept of intentionality. Fourth, in follow-up experiments one might collect data for additional variables such as legal culpability or deserved punishment. This would shed light on whether outcome severity does in fact have the assumed effect on punishment (as data for laypeople reported by Cushman (2008) suggests) and whether the impact is indeed mediated by choice of *mens rea* type. Fifth, while we have focused exclusively on intentionality, the problematic extends to other types of *mens rea* that are commonly distinguished in criminal law (cf. US Model Penal Code, section 2.02). Given that the Knobe effect, among laypeople, can also be found for epistemic state ascriptions, jurors and judges might be increasingly likely to find that an agent acted *knowingly* (cf. MPC 2.02 2b), or *should have been aware of a high probability* of the outcome (negligence, cf. MPC 2.02 2d), the more severe the harm done (cf. MacLeod (2016); Kneer (in preparation); Kneer and Machery (in preparation)). Future research should investigate whether this is indeed the case, and it should also

explore whether professional judges are susceptible to the reported effects across different cultures and legal systems. Lastly, if the severity effect is widely found to arise for experts and laypeople alike, and is considered a bias, strategies must be devised to alleviate its impact so as to guarantee a fair and equal trial for all.

Bibliography

Adams, F. (2015). The Knobe-Effect and the Law. *Methode-Analytic Perspectives, 4*(6), 121-135.

Adams, F., & Steadman, A. (2004). Intentional action in ordinary language: Core concept or pragmatic understanding? *Analysis, 64*(282), 173-181.

Alexander, J. (2016). Philosophical Expertise. In J. Sytsma, & W. Buckwalter (Eds.), *A Companion to Experimental Philosophy* (pp. 555-567): John Wiley & Sons.

Alfano, M., Beebe, J. R., & Robinson, B. (2012). The centrality of belief and reflection in knobe-effect cases. *The Monist, 95*(2), 264-289.

Alicke, M. (1992). Culpable causation. *Journal of personality and social psychology, 63*(3), 368-378.

Alicke, M. (2000). Culpable control and the psychology of blame. *Psychological bulletin, 126*(4), 556-574.

Alicke, M. (2008). Blaming badly. *Journal of Cognition and Culture, 8*(1), 179-186.

Alicke, M., & Rose, D. (2010). Culpable control or moral concepts? *Behavioral and Brain Sciences, 33*(04), 330-331.

Beebe, J. R. (2013). A Knobe effect for belief ascriptions. *Review of Philosophy and Psychology, 4*(2), 235-258.

Beebe, J. R., & Buckwalter, W. (2010). The epistemic side‐effect effect. *Mind & Language, 25*(4), 474-498.

Beebe, J. R., & Jensen, M. (2012). Surprising connections between knowledge and action: The robustness of the epistemic side-effect effect. *Philosophical Psychology, 25*(5), 689-715.

Butler, R. J. (1978). Report on Analysis "Problem" no. 16. *Analysis, 38*(3), 113-114.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.

Cole Wright, J., & Bengson, J. (2009). Asymmetries in judgments of responsibility and intentional action. *Mind & Language, 24*(1), 24-50.

Cova, F. (2016). The Folk Concept of Intentional Action: Empirical approaches. In W. Buckwalter, & J. Sytsma (Eds.), *The Blackwell Companion to Experimental Philosophy*.

Cova, F., Dupoux, E., & Jacob, P. (2012). On doing things intentionally. *Mind & Language, 27*(4), 378-409.

Cova, F., Lantian, A., & Boudesseul, J. (forthcoming). Can the Knobe Effect be Explained Away? *Personality and social psychology bulletin*.

Cova, F., & Naar, H. (2012). Side-effect effect without side effects: the pervasive impact of moral considerations on judgments of intentionality. *Philosophical psychology, 25*(6), 837-854.
Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition, 108*(2), 353-380.
Cushman, F. (2010). Judgments of morality, causation and intention: Assessing the connections. *Unpublished manuscript, Harvard University*.
Cushman, F., & Mele, A. (2008). Intentional Action: Two and a half folk concepts? In J. Knobe, & S. Nichols (Eds.), *Experimental philosophy* (pp. 171-188). Oxford: Oxford University Press.
Dalbauer, N., & Hergovich, A. (2013). Is What is Worse More Likely?—The Probabilistic Explanation of the Epistemic Side-Effect Effect. *Review of Philosophy and Psychology, 4*(4), 639-657.
Duff, A. (1989). Intentions legal and philosophical. *Oxford J. Legal Stud., 9*, 76.
Duff, A. (2015). Intention, Intentional Action and the Law. *Methode-Analytic Perspectives, 4*(6), 136-146.
Feltz, A. The Knobe effect: A brief overview. In *Journal of Mind and Behavior, 2007a*: Citeseer
Feltz, A. (2007b). The Knobe effect: A brief overview. *Journal of Mind and Behavior, 3*(4), 265-277.
Ginther, M. R., Shen, F. X., Bonnie, R. J., Hoffman, M. B., Jones, O. D., Marois, R., et al. (2014). The language of mens rea.
Guglielmo, S., & Malle, B. F. (2010). Can unintended side effects be intentional? Resolving a controversy over intentionality and morality. *Personality and Social Psychology Bulletin*, 0146167210386733.
Hahn, U., & Harris, A. J. (2014). What does it mean to be biased: Motivated reasoning and rationality. *Psychology of learning and motivation, 61*, 41-102.
Harman, G. (1976). Practical reasoning. *The Review of Metaphysics*, 431-463.
Hindriks, F. (2008). Intentional Action and the Praise‐Blame Asymmetry. *The Philosophical Quarterly, 58*(233), 630-641.
Hitchcock, C., & Knobe, J. (2009). Cause and norm. *The Journal of Philosophy, 106*(11), 587-612.
Holton, R. (2010). Norms and the Knobe effect. *Analysis*, anq037.
Kamm, F. M. (1993). Morality, Mortality. Volume I: Death and Whom to Save From It.
Katz, L. (1987). *Bad acts and guilty minds: Conundrums of the criminal law*. Chicago: University of Chicago Press.
Kauppinen, A. (2007). The rise and fall of experimental philosophy. *Philosophical explorations, 10*(2), 95-118.
Kneer, M. (forthcoming). Perspective and Epistemic State Ascriptions. *Review of Philosophy and Psychology*.
Kneer, M. (in preparation). Guilty Minds and Biased Minds.

Kneer, M., & Bourgeois-Gironde, S. (forthcoming). Attribution de mens rea: Données empiriques. In S. Ferrey, & F. G'Sell (Eds.), *Causalité, Responsabilité et Contribution à la Dette*. Brussels: Editions Bruylant.
Kneer, M., & Machery, E. (in preparation). No Luck for Moral Luck.
Knobe, J. (2003a). Intentional action and side effects in ordinary language. *Analysis, 63*(279), 190-194.
Knobe, J. (2003b). Intentional action in folk psychology: An experimental investigation. *Philosophical Psychology, 16*(2), 309-324.
Knobe, J. (2004). Intention, intentional action and moral considerations. *Analysis, 64*(282), 181-187.
Knobe, J. (2006). The concept of intentional action: A case study in the uses of folk psychology. *Philosophical Studies, 130*(2), 203-231.
Knobe, J. (2010a). Action trees and moral judgment. *Topics in Cognitive Science, 2*(3), 555-578.
Knobe, J. (2010b). Person as scientist, person as moralist. *Behavioral and Brain Sciences, 33*(04), 315-329.
Knobe, J., & Burra, A. (2006). The folk concepts of intention and intentional action: A cross-cultural study. *Journal of Cognition and Culture, 6*(1), 113-132.
Knobe, J., & Fraser, B. (2008). Causal judgment and moral judgment: Two experiments. *Moral psychology, 2*, 441-448.
Kobick, J. (2010). Discriminatory Intent Reconsidered: Folk Concepts of Intentionality and Equal Protection Jurisprudence. *Harv. CR-CLL Rev., 45*, 517.
Kobick, J., & Knobe, J. (2009). Interpreting intent: How research on folk judgments of intentionality can inform statutory analysis. *Brook. L. Rev., 75*, 409.
Lacey, N. (1993). A clear concept of intention: elusive or illusory? *The Modern Law Review, 56*(5), 621-642.
Lanteri, A. (2012). Three-and-a-half folk concepts of intentional action. *Philosophical Studies, 158*(1), 17-30.
Leslie, A. M., Knobe, J., & Cohen, A. (2006). Acting intentionally and the side-effect effect theory of mind and moral judgment. *Psychological Science, 17*(5), 421-427.
Levinson, J. D. (2005). Mentally misguided: How state of mind inquiries ignore psychological reality and overlook cultural differences. *Howard LJ, 49*, 1.
Lombrozo, T., & Uttich, K. (2010). Putting normativity in its proper place. *Behavioral and Brain Sciences, 33*(04), 344-345.
Lowe, E. J. (1978). Neither intentional nor unintentional. *Analysis, 38*(3), 117-118.
Ludwig, K. (2007). The epistemology of thought experiments: First person versus third person approaches. *Midwest Studies in Philosophy, 31*(1), 128-159.
Machery, E. (2008). The folk concept of intentional action: Philosophical and experimental issues. *Mind & Language, 23*(2), 165-189.
Machery, E. (2012). Expertise and intuitions about reference. *THEORIA. Revista de Teoría, Historia y Fundamentos de la Ciencia, 27*(1), 37-54.

MacLeod, J. A. (2016). Belief States in Criminal Law. *Oklahoma Law Review*, 497-554.
Malle, B. F., & Knobe, J. (1997). The folk concept of intentionality. *Journal of Experimental Social Psychology, 33*(2), 101-121.
Malle, B. F., & Knobe, J. (2001). The distinction between desire and intention: A folk-conceptual analysis. *Intentions and intentionality: Foundations of social cognition*, 45-67.
Malle, B. F., & Nelson, S. E. (2003). Judging mens rea: The tension between folk concepts and legal concepts of intentionality. *Behavioral sciences & the law, 21*(5), 563-580.
Mele, A. (1992). *Springs of action: Understanding intentional behavior*: Oxford University Press.
Mele, A., & Cushman, F. (2007). Intentional action, folk judgments, and stories: Sorting things out. *Midwest Studies in Philosophy, 31*(1), 184-201.
Moore, M. S. (2011). Intention as a marker of moral culpability and legal punishability. In A. Duff, & S. P. Green (Eds.), *Philosophical Foundations of Criminal Law*. Oxford: Oxford University Press.
Nadelhoffer, T. (2004a). Blame, Badness, and Intentional Action: A Reply to Knobe and Mendlow. *Journal of Theoretical and Philosophical Psychology, 24*(2), 259-269.
Nadelhoffer, T. (2004b). On praise, side effects, and folk ascriptions of intentionality. *Journal of Theoretical and Philosophical Psychology, 24*(2), 196-213.
Nadelhoffer, T. (2006). Bad acts, blameworthy agents, and intentional actions: Some problems for juror impartiality. *Philosophical Explorations, 9*(2), 203-219.
Nichols, S., & Ulatowski, J. (2007). Intuitions and individual differences: The Knobe effect revisited. *Mind & Language, 22*(4), 346-365.
Pettit, D., & Knobe, J. (2009). The pervasive impact of moral judgment. *Mind & Language, 24*(5), 586-604.
Phelan, M. T., & Sarkissian, H. (2008). The folk strike back; or, why you didn't do it intentionally, though it was bad and you knew it. *Philosophical Studies, 138*(2), 291-298.
Robinson, B., Stey, P., & Alfano, M. (2015). Reversing the side-effect effect: the power of salient norms. *Philosophical Studies, 172*(1), 177-206.
Robinson, P. H., & Darley, J. M. (1995). *Justice, liability, and blame: Community views and the criminal law*. Boulder, Colorado: Westview Press.
Sauer, H., & Bates, T. (2013). Chairmen, Cocaine, and Car Crashes: The Knobe Effect as an Attribution Error. *The Journal of ethics, 17*(4), 305-330.
Scaife, R., & Webber, J. (2013). Intentional side-effects of action. *Journal of Moral Philosophy, 10*(2), 179-203.
Schulz, E., Cokely, E. T., & Feltz, A. (2011). Persistent bias in expert judgments about free will and moral responsibility: A test of the expertise defense. *Consciousness and Cognition, 20*(4), 1722-1731.

Schwitzgebel, E., & Cushman, F. (2012). Expertise in moral reasoning? Order effects on moral judgment in professional philosophers and non‐philosophers. *Mind & Language, 27*(2), 135-153.
Schwitzgebel, E., & Cushman, F. (2015). Philosophers' biased judgments persist despite training, expertise and reflection. *Cognition, 141*, 127-137.
Severance, L. J., Goodman, J., & Loftus, E. F. (1992). Inferring the criminal mind: toward a bridge between legal doctrine and psychological understanding. *Journal of criminal justice, 20*(2), 107-120.
Shen, F. X., Hoffman, M. B., Jones, O. D., Greene, J. D., & Marois, R. (2011). Sorting guilty minds. *New York University Law Review, 86*.
Sripada, C. S. (2010). The deep self model and asymmetries in folk judgments about intentional action. *Philosophical Studies, 151*(2), 159-176.
Sripada, C. S. (2012). Mental state attributions and the side-effect effect. *Journal of Experimental Social Psychology, 48*(1), 232-238.
Sripada, C. S., & Konrath, S. (2011). Telling more than we can know about intentional action. *Mind & Language, 26*(3), 353-380.
Tannenbaum, D., Ditto, P. H., & Pizarro, D. A. (2007). Different moral values produce different judgments of intentional action. *Unpublished manuscript, University of California-Irvine*.
Tyler, T. R. (2006). *Why people obey the law*: Princeton University Press.
Ulatowski, J. (2008). *How Many Theories of Act Individuation are There?* PhD Dissertation, University of Utah.
Ulatowski, J. (2012). Act individuation: An experimental approach. *Review of Philosophy and Psychology, 3*(2), 249-262.
Uttich, K., & Lombrozo, T. (2010). Norms inform mental state ascriptions: A rational explanation for the side-effect effect. *Cognition, 116*(1), 87-100.
Weinberg, J. M., Gonnerman, C., Buckner, C., & Alexander, J. (2010). Are philosophers expert intuiters? *Philosophical psychology, 23*(3), 331-355.
Williamson, T. (2008). *The philosophy of philosophy*: John Wiley & Sons.
Williamson, T. (2011). Philosophical expertise and the burden of proof. *Metaphilosophy, 42*(3), 215-229.