Making Sense of the Knobe-effect

Praise demands both Intention and Voluntariness

Istvan Zoltan Zardai
Visiting researcher, Keio University

Abstract

The paper defends the idea that when we evaluate whether agents deserve praise or blame for their actions, we evaluate both whether their action was intentional, and whether it was voluntary. This idea can explain an asymmetry in blameworthiness and praiseworthiness: Agents can be blamed if they have acted either intentionally or voluntarily. However, to merit praise we expect agents to have acted both intentionally and voluntarily.

This asymmetry between demands of praise and blame offers an interpretation of the Knobe-effect: in the well-known experiment people blame a company chairman because, although he harmed the environment unintentionally, he did so voluntarily. In turn, praise is withheld, because the chairman did not benefit the environment intentionally. This is a way of rendering the Knobe-effect a rational outcome. It is an advantage of this position, that the distinction between the intentionality and voluntariness of actions can be upheld, whether or not it is the best explanation of the Knobe-effect.

Keywords: action, Knobe-effect, intention, voluntariness, praise

1. One of the uses that philosophy of action serves to other sub-fields of philosophy, as well as beyond philosophy, is to offer clear views of what actions are, and as such of what the objects of our knowledge of actions, judgments of actions, and evaluations of actions are. Once a general view of action is worked out, a sensible view of intentional and voluntary actions can be offered and applied in other fields, hopefully helping to clarify questions concerning responsibility. This paper shows that making a distinction between actions being intentional and actions being voluntary, helps to understand why praise and blame are asymmetrical. I argue that this asymmetry is illustrated by one of the most exciting findings of experimental philosophy, the Knobe-effect.[1]

According to the paper in which Knobe reported his results, people asymmetrically evaluate actions as intentional when the action has known but unintended harmful results, and as unintentional when the action has known but unintended beneficial results. That is, people are likely to claim that agents intentionally bring about harmful consequences of their intended actions and, as the experiment shows, people think agents should be blamed for such consequences. While at the same time they claim that agents unintentionally bring about the beneficial consequences of their intended actions and should not be praised for them. This asymmetry is called the Knobe-effect, and has spurred a rich literature of interpretation and explanations.

The main claim of this paper is that adopting a view of action which distinguishes intentionality and voluntariness helps to interpret the Knobe-effect correctly by shedding light on an underlying asymmetry in the criteria of moral desert. The asymmetry's explanation is that for an agent to deserve praise the agent has to act both intentionally and voluntarily, while to deserve blame it is enough that the agent acts intentionally or voluntarily.[2] Intentional but involuntary actions, and unintentional but voluntary actions do not merit praise,

1 Named after Joshua Knobe who carried out the original experiment. See his 2003.

2 For alternative accounts of the asymmetry of praiseworthiness and blameworthiness see for example Susan Wolf's 1980 and Dana Nelkin's 2011 work.

but can earn blame for the agent. This means that the asymmetry in our evaluations of actions and our resulting blaming and praising of agents highlighted by the Knobe-effect is rational. If this is the true explanation of the experiment's results, then there is no fundamental incoherence in folk-judgments, nor are people swayed by the moral status of outcomes. It is possible of course, that the distinction between the intentionality and voluntariness of actions is a real distinction that needs to be made, and that the criteria of praise and blame are asymmetric in the way presented here, nevertheless when people evaluated the vignettes of Knobe's experiment they were influenced by other considerations or affected by some psychological mechanism they were unaware of. The explanation presented here can then be still a true view of intentional and voluntary actions, and of the criteria of praise and blame, without being an explanation of the experiment. Carrying out experiments to test whether people are relying on the distinction between intentionality and voluntariness would be the topic of a further project and is not among the goals of this paper. The distinction between intentional and voluntary has been drawn by Aristotle, Kant, and other philosophers.[3] A substantial recent attempt to render the distinction explicit can be found in John Hyman's book Action, *Knowledge, and Will* (2015). The paper relies on Hyman's view to introduce the distinction.

2. Knobe's experiment, which yielded the effect named after him, had the following setup: participants were presented with two vignettes, two stories with a single difference. In the first story the chairman of a company is told that the new strategy worked out will benefit the company if implemented, and it will also have the result of benefitting the environment. In the second story the chairman of a company is told that the new strategy worked out will benefit the company if implemented, and it will also harm the environment. In both cases the chairman decides to implement the new strategy, stating that his concern is benefiting the company and he does not care about helping the environment.

Participants are then asked to judge whether the chairman benefitted the environment intentionally in the first scenario, and whether he harmed the environment intentionally in the second. In the original experiment the majority of participants, 77%, judged the chairman to have benefitted the environment unintentionally, and 82% of respondents judged the chairman to have harmed the environment intentionally. Respondents were then asked how much blame or praise they would assign to the chairman. Participants assigned high rates of blame in the harm case, and low levels of praise in the help case. The effect has been replicated several times with the same and different vignettes too. The following is the original text of the harm-case;

> The vice-president of a company went to the chairman of the board and said, 'We are thinking of starting a new program. It will help us increase profits, but it will also harm the environment.' The chairman of the board answered, 'I don't care at all about harming the environment. I just want to make as much profit as I can. Let's start the new program.' They started the new program. Sure enough, the environment was harmed. (Knobe 2003, 191)

and the following is the benefit-case

> The vice-president of a company went to the chairman of the board and said, 'We are thinking of starting a new program. It will help us increase profits, and it will also help the environment.' The chairman of the board answered, 'I don't care at all about helping the environment. I just want to make as much profit as I can. Let's start the new program.' They started the new program. Sure enough, the environment was helped. (Knobe 2003, 191)

As can be seen the only difference between the two stories presented to participants of Knobe's original experiment is the use of the words 'harm' and 'help'.

3. Numerous explanations of what leads to the incoherence have been proposed. Some commentators are of the view that because the responses were asymmetrical, the result shows that the participants' judgments regarding intentionality were incoherent in an irrational way, perhaps swayed by the moral status of the side-effects (Nadelhoffer 2004, Knobe 2006, Malle 2006); or that some relevant difference has been

3 Ursula Coope (2010) and David Charles (2017) discuss Aristotle's distinction and its different versions in the *Eudemian Ethics* and the *Nicomachean Ethics.* Kant makes use of the distinction in his *Groundwork for the Metaphysics of Morals*, see the examples of the shopkeeper, and of the man who has lost all desire to live, but chooses to go on out of a sense of duty. (Kant 1786/2011, 22-25) In both cases there is an important difference between simply doing something one wants and aims at; and between doing something one wants and aims at, while one is aware that one could do otherwise and nevertheless acting this way out of respect for the moral law. Kant's view is of course different from Aristotle: Aristotle stresses the ability to judge well (recognise what there is reason to do), and develop the habit to act accordingly; while Kant emphasises the role of choosing the right thing for the right reason (out of respect for the moral law). The connections between choice, obligation, and compulsion are complex; compare Hyman 2015, 92-96.

detected by them, which is not apparent from the text and the setup of the experiment, and that difference explains the asymmetry in the evaluations. For example, Shaun Nichols and Joe Ulatowski proposed that there are two conceptions of intention at play, one is motivation sensitive – whether the action was intentional depends on whether the agent wanted to do it – and one is knowledge sensitive – whether the action was intentional depends on whether the agent knew that he is doing it. (Nichols and Ulatowski 2007) Frank Hindriks suggested that the notion of intentional action has a normative component. (Hindriks 2008) Knobe himself (2006) proposed that the folk concept of intentional action is responsive to the moral nature of side-effects. If I understand him correctly, this does not mean that philosophers are wrong about the way they use intention, however, it does mean that philosophers' conception of intention is a technical one and can be used for other things than the everyday conception of intention. This would be a pluralism about conceptions of intention.[4]

Most researchers engaging with the issue attempted to explain why the chairman was evaluated by participants as having acted intentionally and blamed accordingly. This approach relies on the premise that since the chairman did not have a direct intention to harm the environment, he should not be evaluated as having intentionally harmed it and should not be blamed for this. Commentators taking this track found the responses of participants to the help condition less puzzling. That is, the majority of philosophers and psychologists seem to have endorsed the idea that the chairman did not deserve either blame, or praise, and that while the chairman decided to implement the new program intentionally, harming the environment and helping the environment were results of this.

4. My main claim is that from a philosophical perspective it makes sense that people evaluated the two cases asymmetrically. This is so, because people deserve to be blamed for what they did both when they do something intentionally (but involuntarily), when they do something voluntarily (but unintentionally), and when they do something intentionally and voluntarily.[5] The chairman harmed the environment unintentionally but voluntarily. However, people only deserve praise for their actions when they do them both intentionally and voluntarily. Since the chairman helped the environment voluntarily but not intentionally, he doesn't deserve praise for helping the environment. This distinction can render the responses of participants rational and coherent. The interpretations which propose that there are different conceptions of intention – and claim that folks use a different conception from the philosophical ones in a rational way – and the ones which claim that the folk is simply misguided in their application of their conception of intention (which is similar to that of philosophers') due to the moral quality of the actions might still be correct. Still, it is important that, understood in the way I propose here, the experiment does not create problems for philosophical usages of intention, that is, even if there are folk conceptions of intention (distinct from philosophical ones) that make the experiment's outcome rational and coherent it does not mean that those are the only ones that can make sense of it, or that philosophers have to use those same conceptions. So, what I provide in the following pages is mainly a philosopher's explanation of why I think the outcome of the experiment makes sense and why it is correct to assign moral responsibility asymmetrically in cases of praise and in cases of blame.[6]

If we make the distinction between intentionality and voluntariness, and accept that there is an asymmetry in the criteria of praise and blame, then it can be understood why people blame the chairman in the harm case, and why they deny him praise in the benefit-case: participants judged that while the chairman did not harm the environment intentionally, he harmed it voluntarily, and this is grounds enough for blaming him. If it was clear for participants that the chairman deserves blame, they might have expressed this in replying that he acted intentionally. That is, their judgment that the chairman harmed the environment intentionally expresses their view that he deserves blame because he harmed the environment voluntarily.

We can represent Knobe's results and his interpretation of them in this way

	Harm case	Benefit case
Participants' moral judgment	Bad	Good

4 For a similar kind of pluralism about conceptions of action see Sandis 2012.

5 People deserve blame for some of their unintentional and involuntary doings and for some omissions too, since in certain cases it is their fault that they were not in a position to be able to recognize that they were doing something wrong, or that they had to do something, or that they would be forced to do something. Arguably some of these cases are cases of voluntary passivity; see Aquinas *ST* first part of the second part, q. 6., a. 8; also Hyman 2015, 79-81.

6 Merely making the distinction between intentionality and voluntariness does not commit us to any position regarding free will, determinism, or retribution. The distinction can be endorsed by compatibilists and incompatibilists alike. Regarding retribution and therapy, the view does claim that there are clear criteria of blame, but how we should treat those who are blameworthy is a further question.

Participants' evaluation	Harm intentional	Help unintentional
Assigning	Blame	No praise

Figure 1 ***Knobe's interpretation of his results***

That is, Knobe understands people to first judge agents' actions and their outcomes morally, and their usage of 'intentional' and 'unintentional' to track the moral evaluation. Judging the action intentional and unintentional respectively is coherent with whether the agent is deemed to deserve blame or praise for what they did. Judging the actions to be morally bad or good also has the role of justifying blame and praise, rather than making this dependent on independent criteria of intentionality and unintentionality.

I propose the following way to understand the responses of participants, based on the distinction between intentionality and voluntariness

	Harm case	Benefit case
Response-I	Harm unintentional	Help unintentional
Response-V	Harm voluntary	Help voluntary
Morally	Blame	No praise

Figure 2 ***Interpretation of Knobe's experiment after introducing i) the distinction between voluntary and intentional, and ii) the asymmetry of praise and blame***

As can be seen, if we endorse the distinction between the intentionality and voluntariness of actions we can make sense of the judgments without taking participants to base their responses on a preliminary moral judgment of the action, and using 'intentional' or 'unintentional' to express their moral evaluation of the actions. On this view respondents do look for some criteria independent of the moral quality of the actions and their results, namely, they evaluate how agents related to their actions and their results: did they bring them about intentionally and voluntarily? Participants evaluate the action for these two criteria and then assign their blame and praise accordingly. This evaluation makes sense if we accept that it is harder to earn praise than to deserve blame.

The asymmetry arises because to deserve blame it is enough that an agent does something voluntarily but not intentionally (and also the reverse); however, to deserve praise for doing something it is not enough that an agent brings about some good outcome voluntarily, they also have to do so intentionally. In this case the chairman does not deserve praise for helping the environment because, while he did help the environment voluntarily in that he *knew* that starting the new programme will help the environment (he wasn't ignorant), and he did choose to start the programme *without* being subject to *duress* to do so (he wasn't coerced), he did *not* do so *with the goal* of helping the environment. The chairman had not been motivated to help the environment, it wasn't his goal to help the environment, and he didn't do so in knowledge of aiming at doing so; furthermore, he ignored considerations about the environment when making his decision. Praise is then pre-empted by two factors: one is, that the chairman did not intentionally help the environment, and when one does something good unintentionally one does not deserve praise or recognition for it, since one was not motivated by wanting to do a specific good thing or aiming at the good.

The other factor is that while the chairman's attention was called by the vice-president of the company to the fact that the new programme will help the environment, he nevertheless chose to ignore considerations about the environment when choosing what to do. It is reasonable to assume – and it was reasonable probably to do so on the part of the participants of the experiment – that someone in an influential leadership position is a responsible and well-informed enough person to know about the weight that environmental issues carry these days. Choosing to ignore considerations about harming or helping the environment can indicate unacceptable ignorance about their importance, or a character flaw of not caring enough about issues of great public importance. Either way, it is clear that the chairman does not deserve praise for helping the environment, even though he can be said to have helped the environment voluntarily since he was not coerced or under duress to do so. If anything, he could be blamed for his ignorance or flawed character and lack of values.

5. In spelling out what are conditions of doing something intentionally and when someone is acting voluntarily, I rely on the work of John Hyman. Hyman's ideas are an ideal departure point because they take into account the most important work on intention in the last 70 years – Elizabeth Anscombe's and Donald Davidson's views – and build on these. Hyman's view also tries to rectify a regrettable historical development in anglophone philosophy, namely the conflation of intentionality and voluntariness. There are two separate ways in which agents relate to their actions and these have been run together: voluntariness has been misinterpreted as simply a subset or aspect of intentional actions. In the free will debates voluntariness has been conflated and confused with freedom from determinism. We can distinguish between these two aspects of actions in the following way

Doing something intentionally

An agent does x intentionally if they have a

> motivation to do *x* and their doing so manifests this aim (realises the content of their desire which causes their action).[7]
>
> *Doing something voluntarily*
>
> An agent does *x* voluntarily if they do not do it out of ignorance or compulsion. (Hyman 2015, 77) [8]

Agents act voluntarily if they could have chosen otherwise, and were neither acting under ignorance or under compulsion (duress). That is, voluntariness is a negative concept: it is defined 1 in terms of the absence of circumstances which rob agents from their freedom in the sense that they have to do (or undergo) what another agent forces them to do (or suffer), and 2 in terms of the agents being aware of the relevant moral aspects of their actions (or what they undergo), their rights and obligations, and so on.[9] Involuntariness is not opposed to metaphysical determination, but to duress by another person or organization, and by ignorance, that is, by being unaware of what one is doing. Just because someone is able to choose not to comply with coercion, say in a case when one is threatened by a robber holding a gun, it does not mean that if one deliberately, coolly, and rationally complies, then one did so voluntarily. Such actions are involuntary, even if they are intentional. Complying with the robber's demand to hand over my wallet can be intentional – it realises my desire to survive the encounter unharmed and aims at bringing this about – while being involuntary, since I wouldn't hand over my wallet to a stranger unless I would be forced by them. This kind of intentional submission to a threat is non-consensual. (Hyman 2015, 90-91)

Hyman defends a notion of intention that is causal *and* dispositional. He writes that

> In sum, an explanation of an intentional act that refers to the desire the act expressed or to the intention with which it was done is both causal and teleological. It is causal because it refers to a disposition, and it is teleological because the kind of disposition it refers to is a disposition to pursue an aim, in other words, a disposition that is manifested in goal-directed behaviour. (Hyman 2015, 130)

Hyman argues in detail for a view of intention that combines key insights of Wittgenstein, Anscombe and other defenders of dispositional or non-causal explanations, with the stronger points of causal, Davidsonian views. In effect, he proposes that Wittgenstein, Anscombe, and others following them made a mistake when denying that dispositions can provide causal explanations; whereas Davidsonian views of action – Humean Theories, or causal theories of action – miss the point that desires are dispositions, the content of which is an aim, and this aim is expressed in actions, and hence intentional action has a teleological structure. (Hyman 2015, chapter 5) We do not need to enter into the details of Hyman's arguments here. While I think that his position is an improvement on causal theories of action, any view would do for our present purposes that could capture the point that the intentionality of actions tracks whether they are expressions of an agent's desire to achieve an end knowingly. Hyman's view does this explicitly, and this way we get a very clear distinction between the intentionality and voluntariness of actions: the intentionality of actions concerns the agent's desires and their relation to the action – is the action *actually* caused by a desire of an agent to attain an aim –, whereas the voluntariness of the action depends on whether or not the action is done in the *absence* of certain causes (threats and other forms of duress, or obligations that the agent does not want but has to follow, or lack of relevant information about what they are bringing about).

Given these definitions, one could say that the

---

7 For the purposes of the paper, I'm leaving out some details of Hyman's position which are relevant to the debates he considers, but do not make a difference to the proposed understanding of the Knobe effect. For his full view see his 2015; esp. chapter 4 regarding voluntariness, and chapter 5 on intention.

8 Note the similarity of this notion of voluntariness to Aristotle's, worked out in the *Nicomachean Ethics* III 1, 5, and V 8, and the *Eudemian Ethics* II 6-9. As Ursula Coope summarises it, according to Aristotle "**1** an action is not voluntary if it is forced (1110a 1ff.); and **2** an action is not voluntary if it is done in ignorance of the particular circumstances of the action (1110b18ff)". (Coope 2010, 439) Although, the issue is less than clear, since, as David Charles (2017, 10-11) argues, in some respects Aristotle's usage of voluntary seems to resemble our ideas about intentionality, especially in discussions in the *Nicomachean Ethics*; see the example of the captain who, in order to save his ship from sinking, under extreme threat decides to throw overboard some of his cargo. Aristotle talks here of a mixed case, that is partly voluntary since the agent had the ability to choose in a literal sense, but was under compulsion. As Charles notes, this is closer to our notion of intentional, than to voluntary. (Charles 2017, 13-4) Hyman agrees with Charles on this point, and supports the *Eudemian* view. See Hyman 2015, 84-87, including footnotes 24 and 29.

9 Hyman follows Aquinas in recognising the possibility of voluntary passivity: cases when someone undergoes something voluntarily. An example of this would be when one consents to an operation. During the operation one is not active as an agent, but one has consented to being a patient, hence one undergoes the surgery voluntarily. On the connection between consent and the relevant notion of choice to voluntariness see Hyman 2015, 88-91.

chairman did not choose whether to harm or help the environment; he simply chose whether to implement a profitable new programme or not. But this is not true, since the chairman at a minimum chose to put the environmental considerations aside, that is, to ignore considering them in making up his mind about starting the new programme, without being in ignorance about the existence of environmental reasons. He did not necessarily choose to judge the specific environmental harm or benefit caused by adopting the new program, but he certainly judged that in general environmental considerations can be discounted when making such decisions. And he did so while fully informed about the availability of the option of taking the relevant information into account, and also without being under compulsion to ignore it; that is, he did so voluntarily. [10]

6. Some people, among them Knobe (Knobe 2006), suggested that the result indicates that people categorise actions as intentional or unintentional based on the moral status of the results of the actions. This would be an interesting – and for many, troubling – finding. It would show that the status of an action as 'intentional' or 'unintentional' depends not, or not only, on psychological criteria, but on moral ones. As it stands, I don't think this the case, nor that ordinary people would be wrong when they allocate blame to the chairman. The relevant difference that accounts for this is voluntariness.

Knobe's result was evaluated as confusing by many, because it either rendered intention, at least partially, a moral, normative notion. The current ruling view in philosophy of action is however, that the intentionality of actions depends on the psychology of the agent before and during acting. Did the agent want to do the act? Did the agent have a plan to do it? Did this intention play a role in the agent's reasoning before and/or during acting? Did the agent act for reasons they endorsed? Did the agent aim at acting in that way/something realised or achieved by the action? These are the questions which are asked to elucidate in most cases when intentionality is in question. If intentionality is the most prominent feature of actions, then, while it does not follow necessarily, it is a plausible view that praise and blame should be allocated to people depending on whether or not they acted intentionally or not, at least in most cases. This is why the asymmetry in Knobe's result is for many not only perplexing, but also troubling. If it were a correct depiction of what participants thought and how people in general evaluate actions, then it would show both incoherence of judgment, or the falsity of the psychological picture and at the same time incoherence in moral judgment. If people get confused by such a simple and clear-cut case when it comes to allocating blame and praise, then surely the folk cannot be trusted in more complex, real-life cases, when more actors, factors, and results have to be taken into account to arrive at a judgment.

The alternative interpretation I present helps to make sense of the asymmetry. The interpretation claims that the participants did not have the option to choose the correct reply, and as a result they evaluated the chairman as having harmed the environment intentionally to express that he should be blamed. That he acted voluntarily is good grounds for blaming him. So, what the second judgment of participants reflects is that the chairman should be blamed for the harm done to the environment but not because he caused it intentionally, but because he caused it voluntarily. This interpretation of the experiment lays to rest the worry about intentionality. If the respondents distinguished between intentionality and voluntariness, and this is what the asymmetry expresses, then their asymmetrical judgments of the two cases were coherent and rational. The chairman harmed and helped the environment unintentionally; however, he did both voluntarily. This is good grounds for blame, but not enough for praise.

7. Some readers might worry that the distinction I introduce is very technical, and ordinary people would not be able to trace it. I think this is unfounded. Anyone can understand the difference between someone wanting to help do voluntary work to help their neighbourhood – say working on restoring houses after a natural disaster – therefore doing the job, and someone else, let's say a prisoner, not wanting to help but being ordered by the court and the prison to perform this job. What is done is good in both cases – fixing a damaged house – but only in the first case is there a good doing by an agent. The first case is one of acting intentionally and voluntarily, while the second is one of acting intentionally but involuntarily. If the prisoner follows the orders and starts working on the damaged house he is doing so intentionally because he has some motivation to do it. By working on the house, he might be aiming at some further thing – reducing his sentence by showing good behavior, avoiding additional penalties – but even then, he intends to work on the house at best as a means to something else. He is not working on it voluntarily because if there would be no legal pressure on him – no coercion – to do so he would not choose to do so. His will to do so is not something that originates in his character or is in line with his values and views, but

10 While humans can be in circumstances under which financial considerations – losing income, getting into trouble at work, and so on – can count as the kind of duress which renders one's actions involuntary, in the case of the chairman it is reasonable to presume that this condition does not hold. An unsurprisingly small number of chairpersons are destitute or in a position *especially* vulnerable to pressure.

something that is forced on him.

It would be interesting to explore in future experiments whether people would find the prisoner's doing of the repair works intentional or not. In case they would, that would confirm the idea that people can make the distinction between voluntary and intentional, and would recognize here that the prisoner had a desire for an aim – say, to shorten his sentence on grounds of good behaviour – and his action expressed this desire, while at the same time, if there would be no duress to express good behaviour in this way and the task would not be compulsory – meaning that not doing it would count as bad behaviour and against the prisoner's aim – then he would not do it.

Furthermore, almost everyone could understand that both of these cases are different from when someone is absentmindedly tapping on the desk while listening to a lecture (unintentionally, but voluntarily), and from lying on the concrete after being pushed down by a commando unit during a protest (which is unintentional and involuntary). It might be possible that most people who do not work with such concepts regularly – as philosophers, judges, lawyers – would not express the difference by using the words 'intentional', 'unintentional', 'voluntary', and 'involuntary.' Nevertheless, they might be able to understand and track the distinctions, no matter how they would express them. People often use more specific expressions that combine multiple evaluations of an action or an agent, say when we describe someone as a person who relishes attention. Such a remark can explain why they are a good speaker, it can be a compliment in the right context meaning that the person does well in public performances, it can contain an admonition to the effect that perhaps the person cares too much about gaining the attention of others, that their performances are entertaining to others and they have routine in speaking to an audience, and so on.

The idea that sometimes we combine evaluations is lent support by some results of Sverdlik (2004) where he found that it is not merely the moral rightness or wrongness, helpfulness or harmfulness of actions, that influences people's judgment, but also whether the agent regrets what they do. In the examples which participants of Sverdlik's study had to evaluate, Jones has to mow his lawn early in the morning, knowing he will wake up his neighbours. In one scenario, he regrets this, nevertheless since he has to mow the lawn, he does so. In an alternate scenario, Jones does not regret waking up his neighbours, he simply mows the lawn and wakes them up. A higher percentage of respondents evaluated as intentional the case in which Jones did not regret waking up his neighbours, than the one in which he did regret doing so. It has already been pointed out by Aristotle, that regret has a close connection with voluntariness: according to Aristotle, actions done without knowing what one is doing, acting out of a passion – say, when one is drunk – are involuntary, *if they are regretted*. (Charles 2017, 12 and fn. 34; see also Hyman 2015, 98-99) I think a good explanation of why Aristotle thinks so is that lack of regret would show that even if the agent would have known what they are doing, they would have endorsed acting out of the passion manifested in their action, and embraced it as their own. If there is no regret, that shows that neither duress nor ignorance are realised in a way to render the action involuntary, and something the agent is comfortable choosing was done by them. This also indicates that since respondents lack the option of evaluating separately the intentionality and voluntariness of the action, they conflate the two, and what their ratings of intentionality show is possibly their ratings of voluntariness.

My proposed interpretation also avoids the worry about the incoherency or outcome dependency of moral judgments. It does so mainly by offering a more complex picture of judgments about actions. At first sight it might appear that what the interpretation shows is that moral evaluation, and blame and praise, is independent of whether the action was intentional or unintentional, and depends only on whether it was voluntary or involuntary. If this were so, this would lead to an incoherency issue again: it would mean that participants evaluated both the harm and the benefit cases as unintentional but voluntary. If moral status, and blame and praise, depended on voluntariness, then this were a problem, since the judgments regarding voluntariness should be the same, and the asymmetry would be left unexplained. But there is a distinction in the moral judgments and there is an explanation for this, that has to do with moral considerations. What the findings reveal is that the conditions of attributing blame and praise are different. The proposed solution then does not claim that intentionality is irrelevant to praise and blame. Rather to the contrary: one can be blamed for intentional and voluntary actions, for doing something intentionally but involuntarily (like handing over prisoners of war to a cruel detention centre, the commander of which demands this, threatening anyone resisting him with physical retortion), as well as for doing something unintentionally but voluntarily (like drumming on the table while listening to someone's lecture). That is why the chairman is blamed for doing something unintentionally and voluntarily when he harms the environment.

Earning praise is somewhat different: it is more demanding.[11] We don't simply deserve praise for doing

11 At least in one respect. In another, as Dana Nelkin convincingly argues (2011, 39-42), praise demands less: it can be deserved even if agents could not have chosen otherwise or have acted for different reasons, as long as

something intentionally, only if it's also a voluntary action. That the action is voluntary indicates that we did not do it due to compulsion, duress, ignorance, or obligations (at least not due to obligations that we would not follow willingly). That is why voluntary actions expresses what we choose (or is in line with what we would choose). We don't need to make an actual choice to deserve praise, but our actions should be in line with how we would choose if we would do so. Doing something voluntarily but unintentionally does not in most cases deserve praise. In the case of the chairman it is made clear that while the chairman knew about the potential benefits for the environment when approving the new program, he either did not think that was a weighty reason for choosing the policy benefitting the environment – in this case their values seem to be off – or they were negligent and ignored the benefit to the environment taking it to be irrelevant – in this case his negligence can be grounds for evaluating him as unreasonable.

It was possible for the chairman to choose the policy because it was good for the environment, and he did not forego doing so because he was coerced or because he did not know that he could do so. Hence, he counts as voluntarily benefitting the environment, but not as intentionally doing so. It seems then that praise requires agents to do something morally beneficial both intentionally and voluntarily in order to deserve praise for it. If the chairman would care about the environment and would take the benefit for the environment at least as one among several reasons for choosing the policy that could already be enough for him to deserve praise.
The first view of praise and blame, allocating them based on intentions, could be depicted like this

Psychology of the action/results	Intentional	Unintentional
Moral desert	Praise/blame	No praise/no blame

Figure 3 ***Simple View of praise and blame***

This could be called the *Simple View of praise and blame* (SV), which would claim that whether or not someone deserves praise or blame for something depends only on the psychological background of their action and its results, namely on whether or not they acted knowingly for a reason aiming at the attainment of what they wanted, and whether what they deserve is praise or blame depends on the moral status of the action or its result.

The more complex view I outlined in the preceding paragraphs, and which is I think lies behind the participants' judgments is the following

Psychology of the action / *Absence of exculpating factors*	Intentional	Unintentional
Voluntary	Praise possible / Blame possible	Praise impossible / Blame possible
Involuntary	Praise impossible / Blame possible	Praise impossible / Blame possible

Figure 4 ***Voluntary-intentional asymmetrical view of praise and blame***

The asymmetry that we see in people's judgments regarding actions that have unintended negative results and actions which have unintended positive results could be interpreted then along the following way: in the case when the chairman accepts the policy which will have a positive effect on the environment, the information about the environment is irrelevant to him. He chooses based on the company's interests. He had the chance to make a choice in the required sense, however he cannot be said to have acted for the reason that the policy would benefit the environment, he merely did not take that as a reason against the policy. He goes along with the force of the financial considerations. This is not something bad, since he merely ignores a positive result. Hence, he is not blamed, but not praised either, since he didn't do anything praiseworthy; he did not deliberate whether or not to choose the policy and then decided on the basis of the positive impact on the environment to go along with it. So, one might say that in this case he benefited the environment unintentionally, but voluntarily (unless he regrets it, but the example doesn't mention this), and hence he is not praiseworthy.

In the case when the chairman chooses the policy despite the fact that he knows that it will harm the environment the situation is relevantly different. In this case the chairman has the information in the same way as in the previous case, he chooses to ignore it in the same way, and to act solely on financial considerations – the psychological side, the structure of intention, is the same –, however going along solely with the financial considerations while being aware of reasons against them does imply that he could have chosen to do otherwise. And the blame that people assign to the chairman in such cases reflects that ignoring morally relevant considerations and going along with the pressure coming from his job do not exculpate the chairman from blame for morally bad and foreseen results of his action. While

they did the right thing because they recognised that reasons call for acting in that way. Such actions conform to the idea of voluntariness endorsed here which does not require freedom from determinism, simply the absence of potentially exculpating factors.

the chairman can claim involuntariness, people would probably challenge him and deny that the pressure of his job and the company's interest were enough to outweigh choosing differently. Thus, the chairman cannot claim either duress or ignorance.

8. My proposal could be further explored with experimental studies. Here my only aim was to show that interpreting the experiment in light of the distinction between intention and voluntariness enables us to understand why agents blame the chairman – he harmed the environment voluntarily. The distinction also throws light on the underlying asymmetry of the criteria of deserving praise and blame, which then explains why the chairman did not deserve praise, although in both cases his relevant acts were unintentional but voluntary. The goal was to provide a plausible, novel understanding of the experiment's results. The majority of interpretations in the literature maintain that the evaluation of the harm condition is surprising, and that of the help condition is understandable, since the agent seems to cause harm and help the environment unintentionally. This essay tried to show that the Knobe-effect can be interpreted in a coherent way. Such an interpretation makes sense of the reactions of the participants as rational and systematic. The essay works on the assumption that since qualifying actions as intentional or unintentional, voluntary or involuntary, and praising and blaming, are central practices that people rely on every day in manners of all weight – from child raising, through workplace debates, to sorting out legal and political issues – explanations which posit that people are systematically wrong or misled are only fallback options we should resort to if we cannot explain the experiment in other ways.

It is possible that even if we would set up an experiment in which we supply people with training to make their intuitive grasp on what is intentional and what is voluntary, when it comes to judging the agent's bringing about a morally bad or a morally good side-effect, peoples' judgment will be influenced and distorted by the moral quality of the side-effect. That is, if the side-effect is morally bad, they will claim that the action was intentional and voluntary, and if the side-effect was morally good, they'll claim that the action was unintentional and involuntary. This would indicate that peoples' judgments of the intentionality and voluntariness of actions *are* distorted by their moral evaluations of the action and its results. That is, it is possible that peoples' judgments would display the correct asymmetry between praise and blame, but for the wrong reasons. This is of course only true if the role of the conceptions of 'intentional' and 'voluntary' which ordinary English speakers use are not meant to simply express that they want to blame or hold an agent responsible, or that they find the agent's actions morally bad (or good, and then praise or commend them). This is one interesting option to discuss: what if when people talk about intentional and voluntary action they are talking about actions that they want to hold agents responsible for, and blame or praise them for it? In this case I would be wrong to say that their use is distorted. What we would get with the experiments would be perhaps instances of correct usage. In that case, to see how people actually use these concepts and what they mean by them, what their conceptions of intentional and voluntary are, and what their functions are in communication we should study their actual usage more. If this is so, then Gustav Lymer and Olle Blomberg (2019) may be right when they claim that we should rely more on natural exchanges when setting up experiments: even tiny differences in sequencing of information can change peoples' evaluations of cases, hence constructing artificial vignettes will almost always be misleading.

The other possibility is one where we stipulate a correct, objective practice of responsibility, blaming and praising, and would treat intention and voluntariness as criteria which can be objectively – impartially and fairly – characterised. This would fit well my insight that the evaluations should be applied symmetrically to the help and the harm case, and that what should explain the differences in the outcomes is that people focus on blaming and praising so much, that this distorts their judgments (rather than simply the moral character of the side-effects causing such distortions).

9. In this paper I presented the distinction between the intentionality and voluntariness of actions, drawing on recent work by Hyman. I used this distinction to support the idea that the allocation of praise and blame is asymmetric: blame can be deserved if one does something either voluntarily or intentionally. In contrast, earning praise is harder: one has to act both voluntarily and intentionally to deserve praise. I then showed that these two ideas together can offer a new interpretation of the Knobe-effect. The core idea is that the chairman was blamed because he acted voluntarily. He didn't earn praise since he did not benefit the environment intentionally. Whether this is really how people judged the case would need further empirical investigation. Independently of the results of such an investigation, I think the distinction between the intentionality and voluntariness of actions is important and can be defended on its own terms, and the same is true of the idea that praise and blame are asymmetric.

References

St Thomas Aquinas, (1920), *The Summa Theologia of St. Thomas Aquinas* (revised ed.), London: Benzinger Brothers.

Charles, D. (2017), 'Aristotle on Agency', In *Oxford Handbooks Online*, Online publication date May 2017. https://DOI:10.1093/oxfordhb/9780199935314.013.6 Accessed November14, 2019.

Coope, U. (2010), 'Aristotle', In *A Companion to the Philosophy of Action*, Edited by Timothy O'Connor and Constantine Sandis, Singapore: Wiley-Blackwell.

Hindriks, F. (2008), 'Intentional Action and the Praise-Blame Asymmetry', *Philosophical Quarterly 58* (233), 630-641.

Hyman, J. (2015), *Action, Knowledge*, *and Will*, Oxford: Oxford University Press.

Kant, I. (1786/2011), *Groundwork for the Metaphysics of Morals*, Translated by Mary Gregor, edited and revised by Jens Timmermann, New York: Cambridge University Press.

Knobe, J. (2003), 'Intentional Action and Side Effects in Ordinary Language', *Analysis* 63 (3), 190-194.

Knobe, J. (2006), 'The Concept of Intentional Action: A Case Study in Uses of Folk Psychology', *Philosophical Studies* 130, 203-231.

Lymer, G. and Blomberg, O. (2019), 'Experimental Philosophy, Ethnomethodology,and Intentional Action: A Textual Analysis of the Knobe Effect', *Human Studies* 42, 673-694.

Malle, B. (2006), 'Intentionality, Morality, and their Relationship in Human Judgment', *Journal of Cognition and Culture* 6, 87-112.

Nadelhoffer, T. (2004), 'On Praise, Side Effects, and Folk Ascriptions of Intentional Action', *Journal of Theoretical and Philosophical Psychology* 24, 196-213.

Nelkin, D. (2011), *Making Sense of Freedom and Responsibility*, New York: Oxford University Press.

Nichols, S. and Ulatowski, J. (2007), 'Intuitions and Individual Differences: The Knobe-effect Revisited', *Mind and Language 22*, 346-365.

Sandis, C. (2012), *The Things We Do and Why We Do Them*, London: Palgrave Macmillan.

Sverdlik, S. (2004), 'Intentionality and Moral Judgment in Commonsense Thought about Action', *Journal of Theoretical and Philosophical Psychology* 34, 224-236.

Wolf, S. (1980), 'Asymmetrical Freedom', *Journal of Philosophy* 77, 151-166.