Doing, Allowing, Framing: A Case for Moral Heuristics

Camilla F. Colombo

RWTH Aachen University

Abstract

Is doing harm morally worse than allowing it to occur? Our every-day intuitions, supported by a long-standing tradition in moral philosophy, suggest that this is the case. Nonetheless, the study of framing effects and cognitive biases has pointed out that our intuitions over the doing/allowing distinction are far from robust and reliable. This line of research casts doubts over the adequacy of our intuitions in grounding the moral principle "doing is worse than allowing" and seems to downplay the doing/allowing distinction as a cognitive bias or as a byproduct of our flawed reasoning skills. In this paper, I take evidence about framing and biases as a serious threat to the doing/allowing distinction. However, if we aim to explain common-sense morality, we need to account for its widespread use. To keep these two insights together, I build a causal model of the distinction, based on Christopher Hitchcock's self-contained network account, which explains instances of attributions of these two labels. I conclude that the doing/allowing distinction can be better understood as a heuristic: in most cases, it helps us delivering quick moral judgements, but it can also misfire when cases are unfamiliar, underdescribed, or controversial.

Keywords: Doing/Allowing distinction, Moral framing, Causal networks, Heuristics.

1. Introduction

Is doing harm morally worse than allowing it to occur? Our every-day intuitions, supported by a long-standing tradition in moral philosophy, argue that this is the case. After all, drowning a man into a pond and not rescuing a drowning man amount to two different conducts, which we evaluate differently from a moral viewpoint. Nonetheless, more recent studies into cognitive biases, framing effects and moral disagreement have pointed out that, besides clear-cut cases like the pond example, our intuitions over the doing/allowing distinction are far from robust and reliable. In short, a) descriptively equivalent actions can be either characterized as "doings" or as "allowings", depending on the framing of features which should be morally irrelevant; b) an "allowing" action can be perceived by

Argumenta (2024): 1—14
ISSN 2465-2334

DOI 10.14275/2465-2334/20240.col
First published: 1 July 2024

some people as morally worse than a “doing” action. This line of research casts doubts over the adequacy of our intuitions in grounding the moral principle “doing is worse than allowing”, and seems to downplay the doing/allowing distinction as a cognitive bias or as a byproduct of our flawed reasoning skills. If this were the case, there would be nothing morally relevant about the doing/allowing distinction, and we should acknowledge that our moral intuitions lead us astray.

In this paper, I take evidence about moral disagreement, framing, and biases as a serious threat to the doing/allowing distinction. However, I also argue that, if we aim to explain commonsense morality, we need to account for our use of the doing/allowing distinction. This means that, to some extent, we cannot easily dismiss the intuitive judgement that doing is worse than allowing, and that these two conducts are distinct. As Wollard (2015) puts it, giving up the principle “doing is worse than allowing” would have serious consequences on our morality, making it either too permissive or too demanding.

To keep these two insights together, I argue that a suitable way to characterize the doing/allowing distinction is the concept of moral heuristic. In a nutshell, “doing” and “allowing” do capture, in a vast majority of ordinary cases, features of actions which are morally relevant, like the severity of consequences and their likelihood, the intentions of the agent, or the fulfilment of some social norm. These two labels, therefore, which hold a strong intuitive appeal, do amount to a fairly reliable guide to make moral evaluations. Specifically, my argument is that the classification of an action as “doing” or “allowing” constitutes a sort of composite judgement, which takes into account different aspects of the context, and which can be summed up as perceiving the action as “default” or “deviant” in a causal network connecting such action to the harmful outcome. Therefore, this distinction could serve as a shortcut to make moral evaluations in a wide range of cases, while not being morally relevant *per se*.

In order to formalize this intuition, I rely on a long-standing tradition in moral philosophy, which aims to analyse “doing” and “allowing” as two different ways of causing an outcome. *Causal* accounts of the doing/allowing distinction, in fact, suggest alternative models to capture the different types of causation at work in “doing” and “allowing” actions. In this paper, I use Hitchcock (2009)’s “self-contained network” account of causality, which I believe is particularly promising as it takes onboard norm-based consideration, allowing context- and judgement- dependency in attributions of causal relations. Within Hitchcock’s model, every variable involved in the description of a causal event can take either a *default* or a *deviant* value, depending on what it is normal to expect in the given context, or what is the “normal” course of events. For instance, if we were reconstructing the causal relations in the event “I drop a lighted cigarette, a fire starts”, the variable “the oxygen is present in the atmosphere” would take its default value, as this is what we expect given the situation. On the other hand, the novel variable in the context, which does stand up with respect to the normal course of events, would be “I drop a lighted cigarette”, which would thus take its deviant value. Relying on this distinction, with some further technical work, Hitchcock can distinguish among different types of causal networks and causal relations. In this paper, specifically, I define “doing” actions as instances where an outcome counterfactually depends on the agent, within a “self-contained” causal network. “Allowing” actions, on the other hand, describe situations where the outcome counterfactually depends on the agent, within a “non self-contained” causal network. The definition of self-contained and non self-contained networks, as I elaborate in the paper, depends upon the presence of deviant or default variables. This model seems to match our intuitions about what counts as doing (or "action") and what counts as allowing (or "omission"). In the pond example mentioned earlier, for instance, pushing a man into the pond would be modelled as a deviant value in the causal network describing the death of the man by drowning; this conduct, upon my interpretation of Hitchcock's account, would be classified as doing. On the other hand, if I continue jogging while I see a man drowning, we could assign a default variable to this conduct; this, in turn, would acknowledge for our perception of the action as an allowing.

This interpretation also accounts for framing and disagreement, providing an explanation for these phenomena. The identification of "self-contained" causal networks thus depends upon which values are assigned to the variables, and, specifically, which value is set as the "default" for all the variables in the network. This feature reflects my insight that "doing" and "allowing" are defined with reference to the "normal" course of events. The assignment of default values thus incorporates agents' expectations and judgements regarding both descriptive and normative features of the context. We can therefore easily observe moral disagreement and moral framing: doing/allowing classifications may vary depending on what agents think will happen or should happen, and depending on the specific framing people may infer different "normal" courses of events.

In most straightforward, detailed and agreed-upon cases, doing/allowing classifications reliably track other morally relevant considerations such as whether the agent intended the harm or the agent acted violating a standard norm. In particular, doing/allowing classifications may serve in these cases as an efficient moral heuristic, tracking different moral and empirical considerations. On the other hand, when cases are unfamiliar, under-described, or pitch different norms against one another in a fairly extreme way, different doing/allowing classifications are reasonable and justifiable, as different default values are legitimate. In these cases, disagreement is to be expected.

2. The Doing/Allowing Distinction in the Moral Literature

The idea that doing harm and allowing a harm to occur amount to two distinct forms of conduct, with different significance and meaning, strikes us as intuitive and reasonable. We do, in fact, use this distinction in real life for many practical circumstances, such as assigning blame and responsibility and calculating compensations. When forming moral judgements, specifically, at least prior to reflection, we seem to share an overwhelming intuition that doing harm is somehow worse than allowing a harm to occur, and should rank higher in terms of the magnitude of the wrongdoing.

The doing/allowing debate in moral philosophy revolves around two main questions: i) where to draw the line between doings and allowings, and ii) whether this distinction matters morally. That is, whether doing behaviours are descriptively different from allowing behaviours, and whether doing harm is harder to justify than allowing harm. It goes beyond the scope of this paper to explore the different positions and frameworks in the moral literature. It is to be noticed, however, that in the debate concerning question i), we can roughly distinguish between two different approaches to the adequate conceptualisation of the doing/allowing distinction: causal account, which distinguish doing and allowing on the basis of how an agent caused an outcome (see, for instance, Bennett, Woollard, Barry and Overlord), with what I call "norm-based" accounts, which attempt to explain the distinction by appealing to independent moral features (Quinn, Foot, Kagan). With respect to question ii), we can distinguish between "positive" and "negative" (or deflationist) frameworks of the doing/allowing distinction. In the first camp, authors like Philippa Foot, Warren Quinn, Jeff McMahan, Frances Kamm, and Fiona Woollard strive to explain the different meaning that commonsense morality seems to attach to "doings" and "allowings". They take seriously our intuitive judgements about specific cases, and try to build upon them a systematic account of the doing/allowing distinction, which justifies the insight that "doing is worse than allowing". In the second camp, authors like Jonathan Bennett, Shelly Kagan, and James Rachels have however challenged the idea that the doing/allowing distinction, in spite of its central role in everyday moral practice, is morally relevant. They argue that the different significance we attach to doings and allowings is not justified after all, either because this distinction is grounded in morally irrelevant features or because it disappears upon careful analysis. Bennett's influential investigation of the distinction, for instance, claims that doing and allowing merely track different ways in which an agent is related to the harmful outcome, but that such features are heavily context-dependent and non morally significant.

The deflationist thesis upheld in the second camp, nonetheless, is at a hard spot both in a) explaining the persistence and strength of the intuitions underlying the doing/allowing distinction, and b) in building a sensible and coherent moral theory once the principle "doing harm is worse than allowing the same harm to occur" is exposed as illegitimate. As Woolard argues, "If there is no moral difference between doing and allowing, then morality must either be far more permissive than we generally suppose—permitting us to kill to protect our personal projects—or far more demanding—requiring constant sacrifice from us to save the lives of others". In short, even if we take the deflationist arguments seriously, the doing/allowing distinction appears to be a principle we would like our moral theories to uphold. The issue of the moral relevance of the distinction, in conclusion, remains of crucial interest for moral philosophers, and the task of providing a coherent justification for this principle appears to be inescapable.

3. Moral Disagreement and Moral Framing

Starting from the late 70ies, research in behavioural economics, cognitive sciences, neurosciences, and psychology has started to empirically investigate moral principles and moral intuitions, including the doing/allowing distinction. The two main results of these areas of research are the so-called phenomena of moral disagreement and moral framing. Roughly, we define moral disagreement on the doing/allowing distinction people disagree over the fact that the "doing" action is morally worse than the "allowing" action. A famous example in the literature is the Smith and Jones example devised by Rachels (1975: 78–80). In this made-up example, two cousins, Smith and Jones, both have the intention to kill their uncle in order to inherit a large sum of money. Smith gets in the bathroom while the uncle is getting a bath, and drowns him to death in order to inherit. Jones, with the same plan in mind, gets in the bathroom and finds the uncle already drowning. He does nothing, he watches him die and inherits the money. Clearly, Smith is doing harm to the uncle, while Jones is allowing the same harm to occur. Nonetheless, people disagree over the fact that Smith's action is more morally objectionable than Jones's. Another influential case is the "starving one's baby" example: an obvious case of allowing harm which nonetheless is often classified as "harming the baby" in people's intuitions and responses.

A "framing effect" is generally said to occur when two descriptions of apparently equivalent decision problems induce systematically different responses and decisions. This widespread phenomenon in choice contexts has been widely investigated in behavioural psychology and decision theory, following the seminal 1979 paper by Daniel Kahneman and Amos Tversky, and has been gathering strong empirical support. "Framing effects" can be caused by a variety of reasoning biases and, as Levin, Schneider and Gaeth (1998) point out, refer to a wide range of phenomena. As Shafir and LeBoeuf (2002) argue, framing effects are also conceived, in the standard narrative, as threatening and challenging to the traditional "rational actor model" and, in general, to the adequacy, reliability and rationality of human cognitive processes. Kahneman and Tversky, in the same fashion, argue that framing effects are particularly problematic for the normative condition of description invariance, which requires that the same decision problem, in terms of expected utility, must be evaluated in the same way by any rational agent.

In the debate over the doing/allowing distinction, moral framing occurs when these effects induce a different classification of the same action as an instance of doing rather than allowing, and different moral judgements of actions so perceived. The most discussed case of framing in this respect is the Asian Flu example. In their 1983 paper "Choices, Values and Frames", Daniel Kahneman and Amos Tversky introduce the concept of a decision frame and outline the tenets of Prospect Theory, which they regard as a model of how agents actually choose. By way of supporting their proposal, they report and analyse different empirical results, among which is the famous "Asian flu" case. This experimental setting divides the subjects into two groups; the first is faced with the following dilemma:

> Your city is threatened by an "Asian flu" that is expected to kill 600 people, and you have to make a choice between these two alternative vaccination programs:
>
> - If Program A is adopted, 200 out of the 600 people will be saved.
> - If Program B is adopted, there is a 2/3 probability that no-one will be saved and 1/3 probability that all 600 people will be saved.
>
> Which program would you choose?

The second group was faced with the very same scenario, but the choice was instead between C and D:

- If Program C is adopted, 400 out of the 600 people will die.
- If Program D is adopted, there is 1/3 probability that no one will die and a 2/3 probability that 600 people will die.

A and C, like B and D, are clearly extensionally equivalent with respect to lives saved, and describe the same vaccination program: "200 people will be saved and 400 will die" (A and C) and "there is 1/3 probability that 600 people will be saved and no one will die and a 2/3 probability that no one will be saved and 600 people will die" (B and D). Therefore, we could reasonably expect that the percentage of people opting for A and C (or for B and D) would be similar in the first and second groups. Nonetheless, experimental findings showed that 72% of subjects in the first group chose Program A but, in the second group, 78% of subjects chose Program

D. Kahneman and Tversky use the Asian flu case, together with five other experimental settings, as representative examples of how decision frames affect agents' behaviours. In particular, in the Asian flu case, they argue that the two different decision frames do not involve different factual descriptions of the world, but rather assume a different reference point as the baseline. In short, Kahneman and Tversky argue that 1) the reference point matters for choice behaviour and 2) people are generally more risk seeking when it comes to avoiding sure losses from a given baseline, and more risk averse when it comes to pursuing gains from a given reference point. In the first decision problem (the choice between A and B), the use of the phrasing "saving" identifies the 200 lives as a gain, thus seemingly setting the reference point at "all 600 people die". With respect to the baseline "everyone dies", choosing program A would amount to a sure gain from the reference point. Plan B, on the other hand, characterises a "bet", as it involves evaluating a risky prospect. Specifically, with respect to the baseline "everyone dies", Plan B could either deliver a bigger gain (all 600 people saved) or simply make no progress at all from the baseline (all 600 die). When it comes to gains, Kahneman and Tversky observe, decision makers tend to be risk averse, and, given the same expected lives saved in A and B, most opt for Plan A, which guarantees a sure gain. In the second decision problem, the different framing of the decision triggers a different evaluation of the vaccination plans. Plan C, indeed, apparently presents the option of 400 people dying as a loss, as it uses the phrasing "die"; this description thus sets the baseline at "all 600 people live". With respect to this reference point, Plan C therefore involves a sure loss. Plan D, again, amounts to a bet, where either losses with respect to the baseline are completely avoided (no one dies) or a bigger loss could occur (all 600 people die). While C and D are expected-lives-saved equivalent, decision makers mostly opt for D, being risk-loving with respect to losses. In conclusion, according to Kahneman and Tversky, different framings select different reference points as the relevant baseline, namely "everyone dies" vs "everyone lives", and this, in turn, induces a different perception of the options as gains rather than losses. Because of the endowment effect, agents would then tend to value the same numbers of lives more when they feel they already "own" them (or they feel they are already secured); therefore, people are supposedly risk seeking when it comes to avoiding losing lives that are framed as losses with respect to the reference point "everyone lives", and risk averse when it comes to saving lives that are framed as gains from the reference point "everyone dies". Consistently, they tend to choose the course of action that involves a chance to completely avoid any loss (plan D over plan C), but are not as eager to take the same risk to save more lives (plan A over plan B).

Tamara Horowitz (1998) makes a further step in linking the baseline sensitivity described by Kahneman and Tversky with the doing/allowing distinction: according to the author, in the first decision problem, the phrasing identifies plan A as doing good, and allowing harm, while in the second decision problem the phrasing induces a classification of plan C as doing harm. Therefore, agents would be more risk taking when it comes to avoid doing harm rather than when it comes to avoid allowing harm to occur, this motivating the preference reversal.

Another famous example of framing effects involves trolley cases. Petrinovich and O'Neill (1995), for instance, analysed people's responses to a trolley case where they are asked to identify with the bystander who could either let the trolley follow its track and run over five people or throw the switch so that the trolley goes to a side track, running over one person (that is, the problem described in Bystander). Respondents were asked to evaluate the two conducts on a 6-point scale from “strongly agree” to “strongly disagree”. In one group, the options in the trolley case were described using the word “kill”—throw the switch and kill one person or let the trolley stay on track and kill five—while the second group worked with questionnaires where the options were described as “saving”—turn the trolley and save five persons or do nothing and save one. Empirical surveys reported that agents were “likely to agree more strongly with almost any statement worded as Save than one worded as Kill”. Specifically, people were more likely to agree, and agreed more strongly, that throwing the switch was permissible, and morally preferable, when this conduct was characterised as Saving. While people still judged that it was permissible to Switch, it seems that they felt more comfortable with and sure of their decisions when the wording was stated in terms of “allowing”.

But which are the implications of such empirical findings? Sinnott-Armstrong (2008) surveys other similar experimental settings; specifically, he argues that empirical data seems to show that moral judgements can also depend on other framing effects besides wording, such as the order in which the examples are presented to the reader. In “Framing Moral Intuitions”, the author concludes that if our intuitions over doing and allowing are shown to be heavily influenced by supposedly non-morally relevant features, like order or phrasing, these moral intuitions are deeply unreliable. In short, the principle “doing is worse than allowing” is nothing more than as the effect of psychological attitudes, idiosyncrasies and reasoning biases, and it is merely built on the flaws of our reasoning skills.

4. Doing and Allowing as Moral Heuristics

The discussions in moral philosophy and in behavioral, cognitive and psychological sciences open a serious theoretical as well as practical dilemma. On the one hand, we have strong reasons, in the lights of building a coherent and sensible ethical theory, to attempt a rigorous justification of the persistent intuition that doing harm is different (and worse) than allowing the same harm to occur. On the other, empirical research tends to explain the doing/allowing distinction as a cognitive bias or, more charitably, as the byproduct of our (flawed) moral reasoning skills. The role of our intuitions about specific cases is thus downplayed, as moral intuitions seem, under closer scrutiny, generally unreliable, controversial and frame-dependent. These two lines of investigation, moreover, appear to be often taken as distinct and do not engage much one with another.

In this paper, I take from the moral philosophy discussion the idea that, if we aim to explain commonsense morality, we need to account for our use of the doing/allowing distinction. This means that, to some extent, we cannot easily dismiss the intuitive judgement that doing harm, all other things being equal, is worse than allowing harm, and that these two conducts are somehow distinct. Nonetheless, I also look at evidence of disagreement and framing effects. From the empirical and experimental tradition, I thus take the idea that our moral intuitions, if not unreliable, might be context- and agent dependent. My own account of the doing/allowing distinction aims to preserve both insights.

4.1 Moral Heuristics

In building this account, I rely on the concept of moral heuristic, as theorised by Cass Sunstein (2005), and on a specific causal account devised by Christopher

Hitchcock (2001). Sunstein defines moral heuristics as shortcuts, rules of thumb, which are used by agents sed to make quick and summative judgement calls for evaluating cases and complex scenarios, without reasoning over principles, or foundational theories. In his 2003 paper, he gives some examples of such heuristics, which can be "punish betrayal of trust" or "do not knowingly cause human death".

My intuition is that our intuitive classifications of doings and allowings might be exactly an example of such heuristics: these classifications provide a "fast and frugal" method to examine and evaluate complex scenarios where different dimensions and features are relevant to our moral judgement. Specifically, following an insight developed by Bennett (1995), I argue that the fact that an action is perceived as doing harm seems to capture cases where something "abnormal" or deviant happens, which appears to be the causal explanation of the harmful upshot, which is somehow a deviation from the standard course of events. On the other hand, when we define a behaviour as allowing harm, this seems to capture cases where the "relevant" cause of the harm is to be found elsewhere, and not in the behaviour of the agent.

To explain this intuition, let us take the two emblematic pond examples illustrated in the Introduction. If I drown someone in the pond, my intervention "stands out" as a full and satisfying explanation of the consequence, the fact that the person dies. The normal course of events, in this scenario, would have been that the person just kept up with her normal activities, and thus my behaviour is "deviant" with reference to this expected course of events. On the other hand, if I do not save the person who is drowning, the fact that I did not jump in the pond does not qualify as a full explanation of the drowning. Moreover, if we consider the scenario where the person is already drowning, the expected course of events is more likely the death of the person.

Before articulating this position more technically, let us examine the implications of this framework for both the moral significance of the doing/allowing distinction and its context- frame- and or person-dependency. In my account, doing/allowing evaluations reflect whether an action and its outcome are perceived as "deviant" with respect to the "normal" course of events (both in the descriptive and normative sense). The doing/allowing distinction, in a vast majority of ordinary cases, is thus a good proxy for features which are morally relevant: in most cases, it tracks whether the consequences are more likely or severity, whether the agent had the intention to harm, whether the agent acted in fulfilment of some social norm, and so on. In ordinary situation, we can thus argue that these labels deliver a composite judgement with respect to different features of the context, and thus amount to a fairly reliable way to make moral evaluations. Most actions intuitively classified as doing harm, in short, do capture instances where an agent had the intention to harm, violated commonly accepted norms, or was the main responsible for serious harmful consequences. In my framework, however, the doing/allowing distinction also relies on a prior definition of what counts as the "normal" course of events, i.e., what we think could or should happen in a specific case. In this sense, people may disagree on normative as well as descriptive features of the context, and depending on the framing, people may infer different "normal" courses of events, this allowing for the phenomena of moral disagreement and moral framing. My further insight is that the more cases are unfamiliar, under-described, or pitch one against the other different norms, the more disagreement is to be expected.

4.2 Hitchcock's "Self-Contained Networks" Account

In this section, I further articulate my account of the doing/allowing distinction by relying on a causal model developed by Christopher Hitchcock. In doing so, my framework fits in the well-established tradition of causal models of the doing/allowing distinction.

Most recent approaches in the literature on causal relations have employed structural equation frameworks to make sense of counterfactuals (Hitchcock 2001, 2007; Woodward 2003; Woodward and Hitchcock 2003, Halpern and Hitchcock 2013). In his 2007 paper, specifically, Hitchcock tackles, amongst other things, the issue of adequately discriminating acts from omissions, and argues that the idea of "self-contained networks" can successfully capture this distinction. While this model shares most of the features of counterfactual accounts, it can also be considered as incorporating many aspects of so-called norm-based accounts, especially in the assessment of the distinction between default and deviant variables. Before discussing this central insight, I briefly sketch Hitchcock's structural equation framework.

First, let a causal model be an ordered pair <V, E>, where V is a set of variables and E is a set of equations among these variables. For simplicity, a variable here can take two values, where one value represents the occurrence, and the other the non-occurrence of a given event, or of a specific version of the event. Let's take this straightforward Assassination example: Alice poison's the victim's drink, and the victim dies. The variables in the story are:

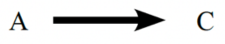
A = 0 if Alice does not poison the drink, 1 if she does;
C = 0 if the victim does not die, 1 if she does.

Hitchcock argues that the counterfactuals we use when discussing the case (in Assassination, "if Alice had not poisoned the drink, the victim wouldn't have died") can be represented by equations among the variables: the variables on the right-hand side of an equation, specifically, work as antecedents of the corresponding counterfactuals, while those on the left work as consequents. In Assassination, the equation describing the causal model is:

$$C = A$$

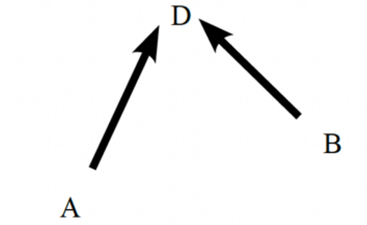
At this point, we can calculate the value of a variable on the left-side of the equation depending on the values taken by the variables on the right-side. For instance, for A = 1 that is, when Alice poisons the drink, we have C = 1, that is, the victim dies. For the equation C = A, we can stipulate that C counterfactually depends on A, because we can compute the value of C fixing the value of A, and the resulting counterfactuals are true: if Alice had not poisoned the drink, the victim would have died; if Alice had not poisoned the drink, the victim wouldn't have died.

For reasons of convenience, Hitchcock suggests that we can represent causal models as graphs, with nodes corresponding to the variables; an arrow from one variable to another represents the fact that the former appears on the right-hand side of an equation with the latter on the left. Hitchcock then defines the former variable as a parent of the latter. For Assassination, we thus have:



where A is a parent of C.

The main strength of this model is that it allows to distinguish between different "types" of causation. To see this point, let us see another example, I call I call Bodyguard: Alice poisons the victim's drink; the victim's bodyguard has an antidote but she does not administer it to the victim. Obviously, the victim wouldn't have died if Alice hadn't poisoned the drink, but she also wouldn't have died had the bodyguard administered the antidote. The causal graph representing this story is the following:



where:

A = 1 if Alice poisons the victim's drink, 0 if otherwise;
B = 1 if Bodyguard administers the antidote, 0 if otherwise;
D = 1 if victim dies, 0 otherwise: and
D = A & not-B.

The difficulty with this case is the one of correctly identifying the different causal impact of the two behaviours. Counterfactually speaking, indeed, Alice's poisoning the drink is causing the death of the victim in exactly the same way the bodyguard's refusing to administer the antidote is: both A and B are thus parents of D. This conclusion, of course, strikes us as intuitively wrong. To solve this problem, Hitchcock defines two alternative mechanisms causation can amount to, each capturing the specific way Alice and the bodyguard are causing the outcome. According to Hitchcock, in Bodyguard, when we read the counterfactual "had Alice not poisoned the drink, the victim wouldn't have died", this appears to be a self-contained story, and Alice's behaviour seems a satisfactory explanation for the victim's death. On the other hand, when we read the counterfactual "had the bodyguard administered the antidote, the victim wouldn't have died" the story is not self-contained or complete: we feel we should know more, as refraining from giving the antidote would not itself and alone bring about the victim's death.

The idea of self-contained or else incomplete causal relationships, relies, in Hitchcock's view, on another distinction, the one between deviant and default values of a variable. The default value of a variable is defined as the value that the variable would take if there was no further information about intervening causes, and the situation were a sort of "self- persisting" system. For instance, in both Assassination and Bodyguard, the default value for C and D is 0, as it is reasonable to expect that, without anyone trying to poison her, the victim would stay alive. A variable which takes a deviant value, on the other hand, amounts to an event that somehow requires an explanation, like the fact that Alice decides to poison the victim's drink. Hitchcock claims that, in the realm of human behaviour, this distinction allows us to identify self-contained versus non-self-contained networks and thus track the act/omission distinction. As should be clear from this explanation, what default values we assign to variables depends on our experience and our judgment; it is not something that we can settle independently of our broader understanding of the situation.

Let's now see in more detail how deviant and default variables can help in distinguishing between self-contained and non-self-contained causal networks. The idea is that we can think of self-contained causal networks as networks providing a “sufficient” explanation of the causal relation at issue. The connection between the drink being poisoned and the victim's death, in this sense, amounts to a satisfactory self-sustaining explanation of the events. On the other hand, a causal network is non-self-contained if it strikes us as incomplete: in short, to explain the occurrence of the outcome, we must appeal to other features which are not included in the network. For instance, the fact that the bodyguard did not administer the antidote is by no means a satisfactory explanation for the death of the victim. According to Hitchcock, we can think that a causal network is self-contained, when, if all the parents of a variable X all take their default value, they cannot cause X to take its deviant value. More intuitively, a causal network

> is self-contained when it is never necessary to leave or augment the network to explain why the variables within the network take the values that they do. When a variable [...] in a self-contained network takes a deviant value, this can be explained in terms of the deviant value of one or more of its parents in the network (Hitchcock 2007: 510).

Let’s be more precise here about what counts, according to Hitchcock, as a causal network. First, Hitchcock introduces the notion of a path as the “set of variables that are all connected by a series of arrows that meet tip to tail”. In Assassination, there is only one path connecting A and C, namely {A, C}. In Bodyguard, {A, D} and {B, D} are the two causal paths connecting A with D and B with D respectively. A causal network connecting variable X with variable Y can then be defined as the set of all variables that feature in paths connecting X to Y. In both these simple examples, the causal networks coincide with the paths: the causal network connecting A with C is {A, C}, while {A, D} and {B, D} are the causal network connecting A with D and B with D. We can now define more formally when a causal network is self-contained versus non-self-contained. Hitchcock provides the following definition, which captures the idea of “sufficient” explanation expressed above:

> Let <V, E> be causal model, and let X, Y ∈ V. Let N ⊆ V be the causal network connecting X to Y in <V, E>. Then the causal network N is self-contained if and only if for all Z in N, if Z has parents in N, then Z takes a default value when all of its parents in N do (and its parents in VÄN take their actual values) (Hitchcock 2009: 412).

Let’s test this formal definition with the Assassination and Bodyguard examples. In Assassination, we can set the default value of C as 0, and the default value of A as 0 as well, since it is not reasonable or natural to expect that someone will poison the drink. The causal network {A, C} connecting A and and C is self-contained: when C takes its default value, its parent A takes its default one as well. More precisely, it is not possible for C to take its default value if its parent takes its deviant one. This matches the intuition that the fact that Alice poisons the drink amounts to a satisfactory and self-sustaining explanation for the death of the victim. What about Bodyguard? Here, the default value of D is set as 0; the default values of A and B are set as 0 as well, as it is not "normal" to expect that Alice will poison the drink, or that someone will administer an antidote.10 The causal network {A, D} is self-contained, as it is not possible for D to take its default value if A takes its deviant one; this, again, matches our intuitions about what counts as a sufficient explanation. {B, D}, on the other hand, is non-self-contained: D takes its default value if B takes its deviant one. This result matches the intuition that B is not a satisfactory explanation of D.

We have now enough elements to formalize my account of the doing/allowing distinction within Hitchcock's model: I define "doing" actions as instances of counterfactual dependence within a self-contained causal network, and "allowing" actions: instances of counterfactual dependence within a non-self-contained causal network. This model, I argue, matches our attributions of doing and allowing. In the pond cases, for example, when I drown the person, we have that "If I had not pushed her (default), the person would not have drowned (default)": the counterfactual is true, and the outcome would have taken its default value when the parent had taken its default one. The network is self-contained and correctly identifies the behaviour as doing harm. In the failing to rescue case, we have that "if I had jumped (deviant), the person would not have drowned (default)": the counterfactual is true, and the outcome can take its default value when its parent takes its deviant one. The network is non-self-contained, this correctly identifying the behaviour as allowing harm.

As a final remark, we must also note that "doing harm is morally worse than allowing the same harm to occur" is not the only nor necessarily the most prominent principle guiding our moral evaluations, or the only moral heuristics in our toolbox. Disagreement over the significance and relative judgements of courses of actions might also depend on other heuristics being involved in the appraisal of a specific case, thus complicating the picture. I suggest that, in this sense, my framework could help isolating the import of the doing/allowing distinction, and avoiding conflating different sources of moral disagreement or moral uncertainty. In other words, it is still possible to hold on the same doing/allowing classifications without agreeing that the doing behaviour is morally worse than the allowing behaviour.

5. Conclusion

In this paper, I have sketched an account of the doing/allowing distinction which both preserves the intuition that this distinction is morally significant and explains why this distinction can be subject to disagreement and frame/context and person dependent. Doing/allowing classifications are morally significant, as "composite judgements" or heuristics, which incorporate considerations that matter for moral evaluation, and can be a shortcut for the complex procedure of case examination described in Section 4. As such, these descriptions of behaviours should be taken seriously and can legitimately serve our everyday moral practice. On the other hand, doing/allowing classifications are also less stable, more controversial, and less clear-cut than some might hope. Different agents, contexts and framings may make salient different considerations, and deliver different doing/allowing descriptions. Unlike cognitive bias theorists, nonetheless, we need not conclude that our moral intuitions lead us completely astray. In many cases, doing/allowing classifications are "robust" and agreed-upon; in these circumstances, we should keep our intuitive judgement that doing harm is morally worse than allowing harm, all other things being equal. In controversial cases, which I argue are often under-described, I suggest that we should look at those things which are not "equal", that is, are frame- or agent-dependent. In this latter sense, my framework, besides providing and explanatory model for our use of the doing/allowing distinction, could also serve as a useful tool for examining disagreement, and to discriminate whether disagreement is due to biases/cognitive aspects or substantial/moral disagreement.

References

Barry, C., Lindauer, M., and Overland, G., 2014. Doing, allowing, and enabling harm: an empirical investigation. *In*: J. Knobe, T. Lombrozo, and S. Nichols, eds. *Oxford studies in experimental philosophy*. Oxford: Oxford University Press, 175–222.

Bennett, J., 1995. *The act itself.* Oxford: Clarendon Press.

Halpern, J.Y. and Hitchcock, C., 2010. Actual causation and the art of modelling. *In*: R. Dechter, H. Geffner, and J.Y. Halpern, eds. *Causality, probability, and heuristics: a tribute to Judea Pearl*. London: College Publications, 383–406.

Hitchcock, C., 2001. The intransitivity of causation revealed in equations and graphs. *The journal of philosophy*, 98, 273–99.

Hitchcock, C., 2007. Prevention, pre-emption and the principle of sufficient reason. *Philosophical review*, 116, 495–532.

Hitchcock, C., 2009. Cause and norm. *The journal of philosophy*, cvi (1), 587–612.

Horowitz, T., 1998. Philosophical intuitions and psychological theory. *Ethics*, 108, 367–85.

Howard-Snyder, F., 2002. Doing vs. allowing harm. *In*: E.N. Zalta, ed. *The Stanford Encyclopedia of Philosophy* (Summer 2002 Edition). https://plato.stanford.edu/archives/sum2002/entries/doing-allowing [Accessed 20 May 2024].

Kagan, S., 1989. *The limits of morality*. Oxford: Oxford University Press.

Kahneman, D. and Tversky, A., 1983. Choice, values and frames. *American psychologist*, 39 (4), 341–350.

Kamm, F., 1996. *Morality, mortality.* Oxford: Oxford University Press.

Kamm, F., 1998. Moral intuitions, cognitive psychology and the harming-versus-not-aiding distinction. *Ethics*, 108 (3), 463–488.

Kamm, F., 2007. *Intricate ethics*. New York: Oxford University Press.

Knobe, J., 2006. The concept of intentional action: a case study in the uses of folk psychology. *Philosophical studies*, 130, 203–231.

List, C. and Gold, N., 2004. Framing as path-dependence. *Economics and philosophy*, 20 (2), 253–277.

McMahan, J., 1993. Killing, letting die and withdrawing aid. *Ethics*, 103, 250–279.

Osman, M., 2014. Dynamic moral judgement. *Psychology*. Online first. http://www.scirp.org/journal/psych

Petrinovich, L. and O'Neill, P., 1996. Influence of wording and framing effects on moral intuitions. *Ethology and sociobiology*, 17, 145–171.

Petrinovich, L., O'Neill, P., and Jorgensen, M., 1993. An empirical study of moral intuitions: toward an evolutionary ethics. *Journal of personality and social psychology*, 64 (3), 467–478.

Rachels, J., 1975. Active and passive Euthanasia. *New England journal of medicine*, 292, 78–86.

Sinnott-Armstrong, W., 2005. Framing moral intuitions. *In*: W. Sinnott-Armstrong, ed., *Moral psychology.* Cambridge, MA: MIT Press, 47-76.

Sinnott-Armstrong, W., Mallon, R., McCoy, T., and Hull, J.G., 2008. Intentions, temporal order and moral judgments. *Mind and language*, 23 (1), 90–106.

Sunstein, C.R., 2003. Moral heuristics. *University of Chicago law & economics*. Online Working Paper No. 180.

Tversky, A. and Thaler, R.H., 1990. Preference reversals. *Journal of economic perspectives*, 4, 201–211.

Woollard, F., 2008. Doing and allowing, threats and sequences. *Pacific philosophical quarterly*, 89, 261–277.

Woollard, F., 2013. If this is my body...: a defence of the doctrine of doing and allowing. *Pacific philosophical quarterly*, 94, 315–341.

Woollard, F., 2015. *Doing and allowing harm*. Oxford: Oxford University Press.