Revus
Revija za ustavno teorijo in filozofijo prava

52 | 2024
Varia

Exclusionary reasons and mental contamination

A challenge for Raz's theory of authority

Giuseppe Rocchè

Electronic version
URL: https://journals.openedition.org/revus/10412
DOI: 10.4000/12nic
ISSN: 1855-7112

Publisher
Klub Revus

Electronic reference
Giuseppe Rocchè, "Exclusionary reasons and mental contamination", *Revus* [Online], 52 | 2024, Online since 11 November 2024, connection on 11 November 2024. URL: http://journals.openedition.org/revus/10412 ; DOI: https://doi.org/10.4000/12nic

Exclusionary reasons and mental contamination

A challenge for Raz's theory of authority

Giuseppe Rocchè

1 Raz's theory of authority and the problem of bias

1 When faced with a certain action, human beings often have reasons for and against that action. Imagine a person is hidden in the bushes with a rifle as a threat approaches. Since her life matters, she has a reason to remain hidden or to flee. But because her people's safety matters too, she has a reason to fire when the threat comes up. This is an example of first-order reasons for action, i.e., reasons to do or not to do a certain action (resting). When a conflict between first-order reasons occurs, the conflict is resolved according to the relative strength of the conflicting reasons.[1] According to Raz, alongside first-order reasons we have second-order reasons, which are any reasons to act on certain other reasons or to refrain from acting on certain other reasons. When a second-order reason is a reason to refrain from acting on certain other reasons, that reason is an exclusionary reason, where refraining from acting on the excluded reasons is compatible both with not doing the act and doing it but not for the excluded reasons. [2] Regardless of whether she has reasons to shoot or not, the woman with the rifle may have reasons for not considering the importance of her life in making her decision.

2 In Raz's thinking, the notion of exclusionary reason is crucial for understanding the concept of legitimate authority and then of legitimate law. According to this approach, in fact, authoritative directives are pre-emptive reasons, which means that they are both a first-order reason for a certain action and a second-order reason to exclude some other reasons –— which may count against that action (or possibly in favour of it). [3] This means that when an authoritative directive prescribes the woman to shoot, the directive is a reason to shoot, but this reason is not just another reason to be added to the other reasons and to be balanced with them. Rather, the directive excludes (displaces, defeats) some of the other reasons, like the relevance of her life.

3 We may call the first-order reasons existing before the directive has been issued "underlying" or "background reasons".[4] According to Raz's *service conception of authority*, whether an authority is legitimate and produces pre-emptive reasons depends on whether people will conform better to the balance of underlying reasons if they try to follow the directive issued by the person or body claiming authority, rather than if they try to follow the balance of underlying reasons directly.[5] The function of authorities is to make people do the right thing. Still, the directive issued by a legitimate authority qua exclusionary reasons must be obeyed even when wrong.[6] For tactical reasons it may be better that the soldier does not shoot at the enemy at that moment. However, the directive supersedes the underlying reasons and the soldier ought to shoot.[7] Since authoritative directives replace some of the background reasons, we may say that they are reasons not to act on the merits of the case at hand — reasons "for doing what you were ordered regardless the balance of reasons".[8] The idea that authoritative directives are pre-emptive reasons is the content of what we may call "the exclusionary model".[9]

4 The exclusionary model is then contrasted with the weighing model. The weighing model does not entirely deprive authoritative directives of normative relevance. The fact that there is an authoritative directive in favour of doing x provides, according to this model, a first-order reason in favour of doing x.[10] The first-order reason may be a very strong reason and may tip the balance, but, unlike in Raz's exclusionary model, it will be a reason that must be balanced with reasons of the opposite sign.

5 One of the rationales of the exclusionary model involves a distrust about human decision-makers.[11] To the extent that well-motivated people are prone to error in weighing reasons for action, it is better that they just try to follow authoritative directives. This distrust in the judgement of moral agents is an argument in favour of the exclusionary model over the weighing model, since in the latter model the agent is still asked to weigh the underlying reasons and authoritative directives as first-order reasons.

6 One of the reasons for distrust in the practical performance of agents lies in their susceptibility to bias. The law and authoritative directives in general can be seen as a tool to overcome psychologically distorting factors.[12] As Gur points out,

> Law's comparative advantage at addressing these problems, it has been argued, lies partly in the fact that its characteristic conditions and modes of decision-making are structurally less susceptible to the above biases than the conditions and modes of decision-making typical of day-to-day individual activity. (Gur 2018: 122)

7 However, the weighing model is incapable of reaping the benefits of having an institutional system already in place.

> The crucial difficulty for the weighing model, however, is that when the weight of those structural advantages [*related to the relevant institution and form of regulation*] is assessed by the actor in situations of bias that made law necessary in the first place, this assessment itself is liable to be biased too. (Gur 2018: 122)

8 Against this background, the exclusionary model promises to better shelter agents from their biases.[13] But the extent to which the exclusionary model actually succeeds in achieving this goal is unclear, and the purpose of this paper is to address this issue.[14] Sections 2 and 3 are dedicated to the premises of the argumentation. In particular, Section 2 delimits the target, clarifying which interpretation of the exclusionary model is under consideration, while Section 3 delimits the type of biases and experiments

challenging the exclusionary model so understood. The following sections constitute the attacks on the exclusionary model, based on experimental evidence. They focus respectively on our inability to intentionally disregard certain factors, in general (Section 4), in the moral sphere (Section 5), and finally in the legal sphere (Section 6). The attack combines a general framework on our inability to intentionally ignore relevant or quasi-relevant information — studies on "mental contamination"; Jonathan Haidt's research on the conflict between moral reasoning and moral intuitions; and studies on motivated reasoning. While the main aim is to support the critique of the exclusionary model, this juxtaposition of empirical research may be useful for other purposes as well. In particular, it could weaken the premises of Haidt's theory about ex post rationalization strengthening its conclusions, and lay the groundwork for further analyses about the relations between law and morality. The article concludes with a discussion of the conclusions reached so far and the prospects for new research that they open up.

2 The exclusion task

9 To shed light on the problem of biases within the exclusionary model, it is useful to start with some passages from Raz's work, in which the nature of obedience to the orders of authority is illustrated. In *The Morality of Freedom* Raz expresses the idea that obeying an authority does not imply any surrender of judgment, meant as a process of *deliberation* about how to behave in the case at hand.

> [N]o surrender of judgment in the sense of refraining from forming a judgment is involved. For there is no objection to people forming their own judgment on any issue they like. (Raz 1986: 40)
> Surely what counts, from the point of view of the person in authority, is not what the subject thinks but how he acts. I do all that the law requires of me if my actions comply with it. There is nothing wrong with my considering the merits of the law or of action in accord with it. (Raz 1986: 39)

10 So Raz distinguishes the sphere of action from the sphere of deliberation.[15] The conceptual premise is that *acting against her judgment* is not surrendering our judgment, and the main idea is that, to obey authoritative directives, we must act on the pre-emptive reasons stemming from the directives, while we are not required to "stop thinking" about the issue. After authoritative directives have been issued, it is no longer up to us to decide how to act on the basis of a balance of reasons, but we are allowed to form our own judgement, as long as this judgement remains unrelated to our behaviour.

11 After the quoted passage, Raz adds:

> Reflection on the merits of actions required by authority is not automatically prohibited by any authoritative directive, *though possibly it could be prohibited by a special directive to that effect*. (italics added) (Raz 1986: 39)

12 The idea seems to be that the concept of authoritative directives does not require one to stop thinking, but that specific authoritative directives may require it. As we shall see, this is a wise clause. However, the overall message seems to be a rather liberal approach to reasoning, deliberation, and its dangers.

> There may be cases where, given that one's judgment is unreliable, there is reason for one not to contemplate the pros and cons if doing so may lead one to act on one's judgment. But given that one's decision or rule puts an end to this danger,

> once it is adopted there is no longer a reason not to reflect, idly, on the merits. (Raz 1989: 1157, n.9)

13 Raz suggests a model of reasoning that may be described as the following: An agent who wants to obey the authority can reflect on the merits of the action — she can analyse the background considerations applicable to case at hand. What is important is that after reaching her conclusion on the basis of her own understanding of background reasons, she somehow neutralizes the output of her autonomous reasoning, by acting on the basis of undefeated reasons only, that is, on the basis of, the authoritative directive. The image may be that of unpacking a certain box, looking inside at its contents, then repacking the box with the contents inside and acting as if one did not know the contents of the box.

14 In other words, the exclusionary model — according to this interpretation, which will be problematized only at the end of this paper — requires us to exclude reasons for actions by performing a mental task that we will call "*the exclusion task*", where by "exclusion task" I mean a conscious and intentional mental action by which the agent neutralises the relevance of certain mental states. The question is whether this model of reasoning is safe, i.e., whether people can compartmentalize their thoughts.[16] Indeed, it would be an empirical challenge to the exclusionary model if people are unable to compartmentalize their thoughts. If the inability to compartmentalize our thoughts is the inability to perform the exclusion task, this means that in many cases we would not be able to live up to the demands of the exclusionary model.[17]

15 At first glance, it might seem that the only relevant issue is the agent's motivation, and then an advocate of the exclusionary model could easily respond that responsiveness to reasons is enhanced on the premise that the agent is motivated to exclude the underlying reasons within the scope of authority. So why should private deliberation be risky if we assume that the agent is motivated to follow the directive when the time for action arrives? In a nutshell, the problem is that obedience is not entirely dependent on motivation, i.e., disobedience can be unintentional. We may be motivated to deprive our judgements, emotions, or intuitions of any normative relevance and believe that we are successfully neutralizing them while we are still under their influence.

16 What conclusion should we draw from our limits in performing the exclusion task? To be clear, what is of interest is not whether the exclusionary model is all things considered preferable to the weighing model, so much as the extent to which the exclusionary model shields us from the problem of biases that concern the weighing model. It may well be that, despite the limitations that will be shown, the exclusionary model remains preferable to the weighing model in this respect. However, the margin of preference towards the exclusionary model over the weighing model could then be reduced, and this could lead us to prefer the weighing model for its other virtues, or to look at other models, or perhaps to interpret the exclusionary model in different ways.

3 Clearing the ground: Awareness and relevance

17 In starting the analysis of the psychological challenge to the exclusionary model, we need to understand which types of biases are more pertinent and which are less pertinent for this challenge. Irrelevant factors can determine our judgment by acting completely behind our backs. A first example is Danziger, Levav, and Avnaim-Pesso's

famous work on the influence of food breaks on parole judges. They found that the 65% approval rate of parole applications after a meal declines steadily in the next two hours, and drops to about 0% before the next break.[18] The anchoring effect provides another useful example. Anchoring describes a phenomenon according to which numbers inserted in a scenario are taken as relevant for the solution of a numeric task, even though they are totally irrelevant.[19] In an unsettling study, for example, judges have been asked to quantify a fine for the violation of a noise-ordinance by a nightclub. Some judges were familiar with a version of the case in which the nightclub was identified as Club 11866 (based on the street address), while others read a version of the story in which the club was identified as Club 58. The number of the identification anchored the judges' response, and those who identified the club with the higher number were more likely to impose a higher fine.[20]

18 Experiments of this kind do not specifically challenge the exclusionary model. For an experiment to challenge the exclusionary model, it must show that people fail to exclude reasons, that is, they fail to successfully complete an exclusion task. But in cases like these, because the disturbing factor is totally irrelevant, and the agent may well be unaware of its dangerous influence, we cannot say that the agent recognizes to be dealing with an exclusion task. Since these cases challenge our faculty to follow both models, they are unable to challenge any one model in particular. The crucial question then is what psychological distortion, and thus what experiments, can challenge our ability to exclude reasons.[21] First, to challenge our ability to perform the exclusion task, we must have experiments focusing on biases that, although identified by the agent, are not corrected; that is, experiments in which the agent tends to be aware of the biasing nature of the situation in which she is operating. So, the first feature of the psychological distortions we are interested in is the *awareness element*.

19 Second, strictly speaking, our ability to perform the exclusion task is challenged if the design of the experiment shows that the biasing information that must be excluded is considered relevant by the agent, i.e., if the information is considered (from her perspective) to be a reason — *relevance element*. If we showed that moral agents can exclude information over facts that counts as reasons,[22] but not information over facts that by no means constitute reasons, we would not have challenged the exclusionary model. We would only know that both the weighing model and the exclusionary model often allow completely irrelevant factors to influence our judgement. But the relevance of the exclusionary model is weakened when it is shown that moral agents are unable to exclude information over facts that count as reasons, since the weighing model, on the contrary, does not stipulate that this information should be excluded.

20 Ideally these two elements would occur together if we are to show the limits of the exclusionary model, and there may be cases where one is present while the other is not. Still, it is also interesting to dwell on the value that each of them has on its own: the inferences we can draw from the existence of one condition in favour of the presence of the other.

21 First of all, there may be a factor relevant for the agent (unlike in the experiment about the anchoring) — notwithstanding her resolution to exclude it —, without the agent having been warned to the need to exclude this factor (like in the anchoring example). In this case the biasing factor is relevant, but its influence is entirely unconscious. These cases do not represent a failure to perform the exclusion task because the agent doesn't even recognize that they are dealing with the exclusion task. However, in some

contexts, it seems reasonable to think — although this must be empirically proven — that the relevance of the factor is evidence of the agent's awareness of its presence and dangerous influence, even in the absence of explicit warning (see V. Moral contamination). Here we would have failures to perform the exclusion task even though the agent has not been explicitly warned about the presence of the biasing factor.

22 The opposite case is that where we have an irrelevant factor for the agent (like in the experiment anchoring example) and an agent who has been warned about the need to exclude it (unlike the anchoring example). I said earlier that to specifically challenge the exclusionary model, "strictly speaking" we need an experiment showing our inability to exclude reasons. Still, experiments showing that agents fail to exclude irrelevant factors when their attention has been drawn to the need to exclude them, are problematic for the exclusionary model in an indirect way. For, if someone she is incapable of excluding irrelevant factors when they are aware that they must exclude them, it is easy to imagine that they are also incapable of excluding what they consider relevant factors. In such a case, empirical evidence to the contrary is always admissible: it may be possible that though we are incapable of excluding irrelevant factors even when we are aware of their influence, we are capable of doing so when they are relevant factors. But it seems reasonable to allocate the burden of proof to the defender of the exclusionary model.

23 Before moving on, one last consideration. So far, I have divided factors, information, and circumstances between relevant and irrelevant — reasons and non-reasons. However, plausibly there are irrelevant factors, information, and circumstances — therefore non-reasons, after all — that somewhat resemble relevant ones: *quasi-reasons*. By "quasi-reasons" I mean factors that have the structure of a reason, in a sense in which random anchoring numbers do not.[23] The fact that my boss was aggressive may be a factor that causes me to challenge him. Postulating that from my point of view, my boss's aggressiveness is not a reason to challenge her, the fact remains that I view this fact as somehow relevant for practical purposes. Phenomenologically speaking, even though for me my boss's aggressiveness is not a reason to challenge her, I may realistically conceive of a person for whom this fact is indeed a reason; while it is extremely hard to imagine someone attaching practical values to random numbers.[24] A bad reason, in this sense, can be a quasi-reason, whereas random numbers never can be. Why distinguish between two types of irrelevant factors, introducing the concept of quasi-reasons? Well, to evaluate the performance of our ability to exclude reasons, one must also consider the case of quasi-reasons as a challenge to the exclusionary model. Just as the burden of proof of the element of awareness can be lightened in the presence of relevant information, so can it be, though perhaps to a lesser extent, in the presence of information pointing to quasi-reasons. Even if for me sexual orientation is not a reason to condemn someone, it is still not the same as random numbers because for me sexual orientation is a salient factor, and I am more likely to be aware that I am faced with an exclusion task.

4 Mental contamination

24 Studies about *mental contamination* provide a useful framework for assessing whether we can successfully complete the exclusion task.

25 Mental contamination may be defined as

> the process whereby a person has an unwanted judgment, emotion, or behavior because of mental processing that is unconscious or uncontrollable. By unwanted, we mean that the person making the judgment would prefer not to be influenced in the way he or she was. (Wilson & Brekke, 1994: 117)

26 In and of itself, this is not an interesting definition, given its level of generality. Moreover, although this expression was previously adopted by some of the most prominent experimental psychologists,[25] it does not enjoy great popularity today. Nevertheless, the idea of contamination serves to emphasize the agent's struggle to debias her thoughts:

> It focuses attention on the difficulty of avoiding many biases. Something that is contaminated is not easily made pure again, which we believe is an apt metaphor for many mental biases. We argue that, because of a lack of awareness of mental processes, the limitations of mental control, and the difficulty of detecting bias, it is often very difficult to avoid or undo mental contamination. (Wilson & Brekke 1994: 117)

27 The authors list four conditions that, if not fulfilled, lead to a contaminated state of mind. First, the agent must be aware that an unwanted mental processing has been triggered. Second, the agent must be motivated to correct the bias. Third, she must be aware of the magnitude and direction of the bias. Finally, she must be able to exercise mental control, adjusting her behaviour.

28 What is relevant for our purpose is that according to this account the detection of the bias and the motivation to correct it are not enough, since for mental contamination to occur it is sufficient to fail in each of the subsequent steps. In particular, the agent may confuse the source of her judgment, believing that she got rid of the bias and that the source of her judgment are undefeated reasons only, whereas the truth is that the bias is still the real source of her judgment — *source confusion*[26]. It is easy to see how the exclusionary model may fall prey to the source confusion. As we have seen, the exclusionary model seems to be based on our ability to compartmentalize our thoughts. The ability Raz seems to be relying on when he says that the obedient agent is not required to surrender her judgment, but only to act on the basis of the authoritative directive. Source confusion is the negation of this ability.

29 The failure of excluding information that is or was somehow relevant for our deliberation is well analysed in studies about *belief perseverance*.[27] According to a popular and intuitive view, which may be traced back to Descartes, understanding and believing are separated mental processes. When a rational agent understands a certain piece of information, it does not mean that she accepts that information as true; on the contrary, she will believe the information only after weighing the arguments for and against its credibility. The Cartesian model of belief formation is contrasted with an opposite model, inspired by Spinoza, according to which understanding a piece of information and believing it are a single mental operation: when a human being understands a piece of information, she begins to take it as true — that is, to believe it — and then decides whether or not to "unbelieve" it. But unbelieving what we believed before is hard because, according to belief perseverance, once the agent has been exposed to a piece of information she integrates the information in her web of beliefs, fabricating explanations revolving around the belief, to the point that if the information is discredited or becomes useless for other reasons, the explanations that the agent has fabricated in the meantime persist, making the belief persist. One case of

mental contamination is provided then by the difficulty of un-believing what was previously believed.

30 Before turning to consider two experiments on belief perseverance it is useful to introduce two notions. Among the reasons to disregard information, it is possible to distinguish between epistemic and extra-epistemic reasons. Epistemic reasons have to do with fact-finding. So, when we have epistemic reasons for ignoring certain information, it means that we have reasons for ignoring information because it is wrong — it does not lead to truth. But while truth certainly has value, it is not the only value, and we may have reasons to ignore information even if it is epistemically valid: extra-epistemic reasons.[28]

31 The first experiment concerns the evaluation of a teacher. [29]Participants were asked to rate a teacher assistant on a scale from one (least nice) to ten (nice). In the first experimental group, some experimenters in disguise gave participants negative information about the teacher. In the second experimental group they gave negative information too, but afterwards they asked participants to disregard the information for *extra-epistemic* reasons, such as "I probably shouldn't have told you those things." In the third experimental group they gave negative information, but afterwards they asked participants to disregard the information for *epistemic* reasons, saying that they were confused and they were referring to another person. Finally, in the control group the participants did not hear any negative information about the teaching assistant. Two things are interesting about this study. First, the evaluations by the first and second group were very similar, and both were very different from the control group. This shows that suppression was totally ineffective in the second group. Second, both the second and third groups had been exposed to the information and were asked to suppress it. However, the rating provided by the third group was statistically higher than the one by the second group. This means that the reasons supporting suppression are relevant to its effectiveness: in this case, epistemic reasons worked better than extra-epistemic reasons.

32 But we should not place too much hope in the reasons given in favour of suppressing information. In a second relevant experiment, subjects were instructed to read some suicide letters. They were informed that some were written by the experimenters and asked to identify the genuine suicide letters. Their performances were then ranked by the experimenter. Some of them performed well, while others poorly. Later, in a second stage, the experimenter revealed that all the letters were fake. In a third stage the experimenter asked the participants to rank their ability to recognize real suicide letters. The results of this experiment showed that those who were evaluated positively at the beginning of the experiment, expressed that they would have performed well even in a real task, while those who were given negative evaluations at the beginning answered that they would have performed poorly.[30] So, even if they had epistemic reasons to ignore the evaluation, the information they acquired persisted in their thoughts.

33 The idea of belief perseverance challenges the exclusionary model insofar as it casts doubt on our ability to exclude reasons. From the agent's perspective, a certain fact looked like a reason for adopting a certain judgment, but even after the agent discovered that the fact in question is not a reason to adopt that judgment, the fact continues to influence their judgment.[31] The fact is no longer a reason, but its ghost continues to influence the agent — a phantom reason.

Finally, it is interesting to highlight some additional aspects of the failure of disregarding information:

> In large part, this belief perseverance resulted from the subjects' tendency to try to explain to themselves why they had performed well or poorly. For example, one subject, told she had done well, stated that she had concluded she was good at evaluating suicide notes because she enjoyed the poetry of Sylvia Plath, who had killed herself. Even though the feedback she had received was discredited, her new beliefs persisted. (Wistrich, Guthrie & Rachlinski 2005: 1268)

34 The quoted passage is relevant because it draws attention to the confidence we have in our ability to disregard information, and to the confabulatory character of the reasons that we put forward to support our judgments. While the subjects' evaluations of their skills were affected by their fake evaluation in the previous stage of the experiment, they believed they were disregarding their fake evaluation and fabricated reasons in support of their self-assessment, which were merely a form of *ex post* rationalization. Summing up the scheme we are interested in and that challenges the exclusionary model, we might represent what we have so far as follows:

> **Mental failure** — (i) The agent is exposed to a distorting factor; (ii) the agent is somehow aware that a distorting factor is in the air; (iii) the agent is honestly motivated to disregard the distorting factor; (iv) the agent believes that she is successfully neutralizing it; (v) but the truth is that the distorting factor is still influencing the agent's decision; (vi) in defending her decision the agent puts forward arguments, but the arguments are a form of ex post rationalization.

5 Moral contamination

35 The experimental examples related to belief perseverance analysed in the previous section are relevant since they show that beliefs of any kind may spoil our judgments. The purpose of this section is to focus on moral beliefs and morality in general: morality as a distorting factor.

36 A clarification needs to be made in this regard. It might seem unusual to consider people's morality as a distorting factor. Certainly, a person may obey moral principles that we consider to be wrong, but this is not the sense in which morality can be said to be a distorting factor. We must remember the definition proposed above: for there to be mental contamination we must have "an unwanted judgment (...). By unwanted, we mean that the person making the judgment would prefer not to be influenced in the way he or she was." In other words, the agent's morality must be an undesirable influence not from the point of view of the external observer, but of the agent himself. We commonly conceive biases as obstacles for the right thing to do, that is for the fulfilment of our moral requirements. How, under this assumption, could a person's morality be the source of a judgment undesirable *to her*? It can be if we imagine the agent's morality as a layered morality, following, for example, Jonathan Haidt's suggestions.[32]

37 In a nutshell, Haidt holds that people's *intuitive* morality is articulated in six moral foundations.[33]

1) The care/harm foundation
2) The fairness foundation (fairness as proportionality)

3) The liberty/oppression foundation
4) The loyalty/betrayal foundation
5) The authority/subversion foundation
6) Sanctity/degradation foundation[34]

38 Along the lines of dual process theories,[35] Haidt believes that our moral practice should be analysed through the distinction between intuition and reasoning. Moral judgment is the result of our intuitions, while reasoning comes after the judgment has been adopted. People are generally led to think that their judgment is the inferential product of reasoning because, while the intuitive part of moral judgment is unconscious, reasoning is a conscious activity, and it is the only part of the process which is experienced by the agent.

39 According to this framework, since our moral judgments are very often causally determined by our intuitions, and moral foundations are the clusters of our intuitions, moral foundations dominate our moral responses. In some cases, our moral intuitions may be seen as a sort of heuristic, anticipating the judgments that we would adopt through the effortful process of moral reasoning. But Haidt is generally interested to another type of situation, in which there is a discrepancy between the moral foundation determining the judgment and the moral principles featuring in our justification.[36] For example, some people, in evaluating a case of incest in which siblings took different contraceptive measures, are driven by the sanctity/degradation foundation to condemn this harmless but offensive conduct. When they are asked to provide reasons for their condemnation, they resort to shaky justifications related to the harm principle, such as the risk of having offspring with some malformation due to endogamy, even though the story clearly stated that the siblings took measure to avoid this risk. Reasoning is a form of *ex post* rationalization. One ingredient of this mismatch between the factor causing the judgment and its justifications is the pressure of social context to conceal some of her moral foundations.[37] But the explanatory centrality of social pressure does not mean that the agent is insincere when she refers to the harm principle to support her judgment: to the contrary, because the real source of the judgment remains unconscious, the agent honestly believes that her condemnation of incest is based on her commitment to the harm principle.[38]

40 The idea of morality as a distorting factor in practical judgment is also at the heart of a tradition of empirical studies that will be taken up in the next section: studies on motivated cognition or motivated reasoning. In motivated reasoning the agent's preferences distort the process of belief formation.[39] Within this general scheme, the agent's morality is one among the relevant preferences capable of orienting her reasoning, leading to a motivated *moral* reasoning.[40] As in the case of Haidt's intuitionism, in the case of motivated reasoning, the agent operates in good faith under "the illusion of objectivity", to use the expression employed by psychologists.[41] One of the differences with Haidt's account that is sometimes emphasized in studies about motivated reasoning, relates to the limits of the agent's ability to form credible justifications for the conclusions she wishes to reach.[42]

41 Through the idea of layered morality, we can imagine unconscious morality as a distorting factor. More specifically we can imagine an internal conflict within the agent between her unconscious morality and her conscious morality (or her non-moral conscious ends) in which the unconscious morality will be a source of mental contamination: moral contamination. That said, for the agent's morality to be a specific

problem for the exclusionary model, we have to imagine that the agent recognizes that she is faced with an exclusion task that she is motivated to complete, where this task is in relevant cases doomed to failure. In other words:

> **Moral failure** — (i) The agent's morality is a distorting factor; (ii) the agent is somehow aware that the issue she is judging triggers her morality; (iii) the agent is honestly motivated to neutralize her distorting morality; (iv) the agent believes that she is successfully neutralizing it; (v) but the truth is that the distorting morality is affecting the agent's judgment; (vi) in defending her decision the agent puts forward arguments, but the arguments are a form of ex post rationalization.

42 The Moral Failure scheme challenges the exclusionary model, and is based on the idea of morality as a distorting factor, an idea developed – as we have seen – by different psychological traditions. Still, there may be relevant differences between the Moral Failure scheme and those theories about practical reasoning. So, it is relevant to see how these theories pose a threat to the exclusionary model. In this work I will limit the analysis to the comparison between moral contamination and Haidt's view, focusing on the relevance and awareness elements.

43 Starting from the relevance element, a direct challenge to the exclusionary model, as we said, would require proof that people are unable to exclude what they believe to be reasons. The relevant question is then whether in Haidt's studies about moral judgment the agent's moral foundations are always a source of reason for action from her point of view. The answer may seem positive, since morality is, one might say, the source of reasons for action *par excellence*. But, as we already know, when we speak of "morality," adopting the approach of psychologists such as Haidt, we mean "morality" in a very broad sense — a layered morality — such that it is not true that all facts to which the agent's "morality" is receptive are reasons for action from the agent's point of view. And when Haidt speaks of "morality" as a distortive factor, he is referring to that part of an agent's unconscious morality that is sensitive to facts that are not practical reasons. The fact that, for example, incest elicits a reaction of disgust is not a reason to condemn incest for Haidt's liberal agent.

44 To challenge the exclusionary model, we can dispense of the relevance element, as long as we have the awareness element. But here there is another problem: Haidt's theory of moral judgment is grounded precisely on the rejection of the awareness element. Indeed, Haidt hypothesizes that moral judgment is the result of intuition, which is equivalent to saying that the agent's moral foundations act unconsciously. It is significant in this regard that Haidt points to intuition and not emotion as the causal factor of judgment precisely because he wants to emphasize how often the mechanism that causes moral judgment is not perceived by the agent.[43] Our morality would catch us off guard, and there would be no moment when we conceive of our task as an exclusion task.

45 Whether moral contamination is a problem for the exclusionary model depends on how frequently the Moral Failure scheme is realized as an alternative to the scheme prefigured by Haidt. This is a question that cannot be resolved at the speculative level but would require experimental confirmation. However, some considerations can still be offered in this regard.

46 The crucial question is whether it is really plausible that most people who find themselves in one of the situations devised by Haidt (situations in which they are offended by a harmless conduct in a social context in which moral condemnation is limited to harmful actions) actually have no inkling that the action is stimulating their moral emotions. While it seems plausible to imagine that the influence of random numbers in anchoring operates entirely outside of our awareness, it does not seem equally plausible that the agent is completely unaware that the incest story is provoking an aversive emotional reaction. It is plausible that the agent often understands that she is disgusted by certain conduct, but at the same time believes that she is neutralizing the impact of disgust on her judgment, and so acting on the basis of undefeated reasons only (those stemming from the harm principle).

47 Relevantly, this eventuality should not be hard to accommodate in Haidt's model. Haidt's most important thesis concerns the confabulatory character of moral reasoning, whereby reasoning does not precede but follows moral judgment. This powerful conclusion is compatible with the Moral Failure scheme: moral reasoning remains confabulatory both in the case where it follows an intuitive judgment, and in the case where it follows a judgment that the agent mistakenly believes she has reached by neutralizing some of her emotions. The mental contamination perspective can be integrated into Haidt's theory, weakening its premises so as to strengthen its conclusions.

48 Before continuing, it is important to make a concluding remark for this section. In this section I have argued that morality might be a problem for the exclusionary model. It may seem that some sort of trickery is hidden in my argument. It may seem that I am saying that we can do without the relevance element because of the awareness element, and then that we can do without the awareness element because of the relevance element. But I do not think that there is such a flaw. And it is here that the idea of a quasi-reason becomes crucial. It is plausible that people, unlike in the case of anchoring with random numbers, realize that some facts (like matters of sex, race, religion, or physical appearance), although irrelevant to the specific task, are generally salient and capable of triggering some unwanted reaction. This ambiguous relevance may lead people to be aware of the danger inherent in the information they have encountered, and so people’s susceptibility to quasi-reasons can be a threat for the exclusionary model of reasons even though quasi-reasons are not reasons, and even though awareness about the exclusion task has not been raised through specific warnings or other measures.

49 Be that as it may, even if studies on practical judgment like Haidt’s were unable to challenge the exclusionary model as long as we remain within pure moral reflection,[44] because we believe that in this context the agent is not sufficiently aware that she is faced with an exclusionary task, or because the excluded considerations are not reasons, a more convincing attack on the exclusionary model can be made once we move to the legal realm. In fact, the law and legal culture often explicitly require the agent, when applying the law, to exclude certain considerations that might seem relevant from a practical point of view.[45]

6 Legal contamination

50 As was mentioned in the opening section, the notion of exclusionary reasons is central for understanding the authority of law. A crucial aspect of the ideal of Rule of Law is that, in Fuller's words, there must be a sort of "congruence between official action and the law", or, following Bruno Celano's analysis, that the law in some sense *predetermines* the decision to be taken in a particular case.[46] For Raz "what the doctrine [of the rule of law] requires is the *subjection* of particular laws to general, open, and stable ones".[47] The basic problem — inherited from the previous discussion — is whether the agent is capable of neutralising her own morality in order to be able to correctly apply the authoritative directive, or whether morality contaminates the application of the law. The scheme of the moral contamination of law, which for simplicity's sake I will call "legal contamination", can be summarised as follows:

> **Legal failure** — (i) The agent's morality is a distorting factor for the correct application of the rule; (ii) the agent is aware that her distorting morality has been triggered; (iii) the agent is honestly motivated to neutralize her distorting morality; (iv) the agent believes that she is successfully neutralizing it; (v) but the truth is that her distorting morality is still influencing the agent's application of the rule; (vi) in defending her application of the rule, the agent puts forward arguments, but the arguments are a form of ex post rationalization.

51 The scheme seems to contrast moral reasons and legal reasons, i.e., the reasons for following the authoritative rule. This approach would contradict the spirit of Raz's theory of authority. For Raz there is a continuity between jurisprudence and moral theory, and the reasons to follow the authoritative directive (instead of acting on the balance of background reasons) are moral reasons too:[48] the conflict between morality and authority must therefore be represented as a conflict between morality and morality.

52 One way to understand this conflict is the idea of a layered morality — introduced in the previous section. But the notion of layered morality requires here further elaboration. While in Haidt the idea of layered morality evokes the distinction between intuition and reasoning, in Raz's theory it points to the distinction between first-order reasons and second-order reasons. Morality is layered in both theories but each of them is focused on different layers. Still, their respective analysis may be fruitfully integrated.

53 We can distinguish three levels of morality. First, there is the unconscious morality that is at odds with the moral principles consciously held by the subject. In Haidt's analysis, the moral foundation of sanctity and degradation would play this role in many peoples' minds. Second, there is the conscious morality that concerns the background reasons. We are in this sphere when we try to do the right thing by balancing the various background reasons. Finally, there is our conscious morality, which recognises the existence of authority and exclusionary reasons. The transition from the second to the third morality is marked by the advent of epistemological and coordination problems, it is fuelled by the dangers of a homemade decision-making process of weighing reasons.[49] Once authorities are recognized, the second morality is, from the point of view of the obedient agent, phenomenologically a kind of "shadow cabinet":

although we know that rules are fundamental and must be obeyed, we often simulate what we would have more reason to do in a world in which there were no legitimate authorities. According to Raz, this simulation — the installation of the shadow cabine is not precluded by the presence of legitimate authority, and this is why authorities do not require us to surrender our judgement. Having said this, in what follows, for expository convenience, the contrast between the first two moralities and the third morality will often be rendered as the contrast between morality (conscious or unconscious) and the law.

54 But before moving on, it must be noted that while the role of the first and third type of morality is univocal, the role of the second is rather ambivalent. The first type of morality may be source of moral contamination, the third morality plays the role of the control system, but the second type of morality for its part can play both roles, sometimes contaminating the decision, sometimes limiting the contamination coming from the first morality.

55 Two different scenarios illustrate the ambivalent character of the middle-level morality. In the simplest scenario the agent's conscious morality considers a fact F1, for example, the belonging to a discriminated minority, as a reason for action A, for example, the granting of some social security benefit. But the agent, as a good citizen or public official, believes she has to abide by legal rules that exclude F1 as a reason for A. In this case, if the agent is driven by her conscious morality, unknowingly or otherwise unintentionally, we would say that she followed her morality rather than authoritative law. *Ex post* rationalization — about the fact that there are reasons for A — would manifest at the level of legal reasoning only, not also at the level of moral reasoning.

56 Additionally, we may imagine a subject who is sensitive to a certain moral foundation to the extent that she believes that fact F2, for example, citizenship, is a reason for A, and who not only adopts the legal rule excluding F2 as a reason for A, but also a conscious moral stance that already excludes F2 as a reason for A. In this case we have two constraints and a source of contamination: the moral foundation may violate the moral constraint and be blocked by legal rules. The Rule of Law would help people following their reflective morality against their unconscious morality, but the moral foundation may violate both constraints. This time the agent hasn't followed her conscious morality instead of the law, rather, the agent's actions have been driven by her unconscious morality instead of by the law and by her conscious morality. *Ex post* rationalization would manifest both at the level of legal and at the level of moral reasoning.

57 Let us then look at two experiments that show that agents are unable to exclude reasons that they believe to be morally relevant or that otherwise stimulate their unconscious morality, when they have to apply legal rules. These two experiments have been chosen because the legal provision or the design of the experiment draw the agent's attention to the exclusion task. To be clear, according to the Razian model, authoritative directives in general are reasons to perform an exclusion task. Still, if we want to prove that people are unable to perform the exclusion task, it is better to focus on cases where the exclusion task is salient because of the peculiarity of the rule involved or it has been rendered salient artificially through the experimental design. This way we can overcome some of the doubts raised at the end of the previous section. Additionally, for the experiments to pose a challenge to the exclusionary model it must

be clear not only that the agent was aware of the exclusion task but also that she was motivated to perform it. As we shall see, while the awareness requirement is easy to meet, this is not the case for the element of motivation.

58 In the first experiment, Wistrich, Guthrie, and Rachlinski wanted to test the efficacy of a "rape-shield" statute, in the deliberative process of real judges. The rape-shield statute is an exclusionary rule that limits the admissibility of information on the chastity and sexual history of the persons involved in the case.[50] The purpose of this provision mixes both extra-epistemic and epistemic concerns. On the one hand, the law wanted to avoid hostile examinations of the alleged victim by the defendant's lawyer. On the other hand, there are epistemic concerns about the admissibility of evidence concerning sexual history. Details of sexual history may be totally irrelevant, or they may be relevant, but in a very uncertain sense. According to some, for instance, promiscuous sexual conduct may suggest that the sexual intercourse was consensual; according to others, promiscuous sexual conduct is a favourable condition for the credibility of a rape report: sexually experienced persons may be more credible, since, in their long history of sexual intercourse, they have not falsely accused their partners of rape.[51]

59 The story depicted in the plot of the experiment is that of a college student who is accused of having sexually assaulted a fellow student during a fraternity party. The complainant had been seen speaking with the accused male at the party before they were seen moving into a room together: apparently the girl was drunk and the boy was helping her walk. The complainant admitted to having had sexual intercourse with the defendant but denied having given consent. Evidence indicates that she immediately contacted the police after the event and that bruises consistent with a sexual assault were found on her body.[52] Judges are asked whether they would convict the accused for sexual assault. But there is a crucial additional detail to take into consideration. In the suppression group, the experimenters added another piece of information. The defendant's attorney tried to introduce the following evidence:

> In his defense, Mr. Geiger is trying to introduce testimony from five other students, 3 male and 2 female, that Ms. Smith had a well-deserved reputation for being sexually promiscuous. This includes one of Ms. Smith's best friends who will testify that before Ms. Smith met her fiancé, she "had trouble remembering what fraternity house she woke up in each Sunday morning." Another witness, a former roommate of Ms. Smith will assert that Ms. Smith "liked to loosen her inhibitions with a few beers too many and then have rough sex with the first guy she saw". (Wistrich, Guthrie, Rachlinski 2005: 1301)

60 It was also added that "the prosecution has moved to exclude such evidence on the ground that it violates Arizona's 'Rape Shield' statute . . . which forbids the introduction of evidence concerning a victim's 'chastity' or 'reputation for chastity' in cases involving sexual assault".[53] The judges were then asked whether the information should have been excluded — as requested by the prosecutor — or admitted.

61 We have three groups of judges to compare. Those in the control condition who never heard the testimony and knew nothing about the sexual history of the alleged victim; those who heard it and found it inadmissible; and those who heard the testimony and admitted it. The interesting result is that while the conviction rate was almost 50% among judges who were not exposed to the testimony, this rate was only 20% among those who heard the testimony but suppressed it. Statistically speaking, there was no difference with judges who read the testimony and admitted it: "In effect, it made no

difference whether the judges who read the inadmissible evidence excluded or admitted it; regardless of their rulings, they relied on it".[54]

62 At first sight the experiment seems to be a promising way of scaling down the advantage of the exclusion model over the weighing model. The way the experiment is designed seems to give an evocative representation of how subjects would behave if, on one hand, they acted on the basis of the balancing of background reasons or on the basis of the weighing model,[55] or, on the other hand, if they acted on the basis of the exclusionary model.

63 The performance of the control group indicates that people who are unaware of the additional testimonies tend to condemn the college student. In contrast, the performance of the judges in the experimental group who admitted the testimonies indicates that people aware of this information tend to acquit. These performances can be seen as a representation of the balancing of background reasons as a decision-making procedure. A "representation", not a proof of the adoption: for all we know, the judges who admitted the evidence could have adopted the exclusionary model, while believing that the law did not require them to exclude the evidence (law is indeterminate, and must be interpreted). But in this case, since they believed that the exclusionary model did not require them to perform an exclusion task, they also relied on the balancing of background reasons. Therefore, we can say that, in this case, people who follow the background reasons tend to acquit.

64 And finally, we have the performance of the judges who knew the testimonies and declared them inadmissible. These judges apparently believed they were facing an exclusion task, and then their performance can be seen as a representation of the exclusionary model. The performance of the exclusion task should lead to the same result reached by the control group, instead it leads to the same result as the people who balanced the background reasons.

65 It seems that for many of the judges who excluded the evidence and convicted the defendant, the evidence unconsciously influenced their judgement. This would be a case where professionals – trained to apply the law – recognise that the law requires them to exclude certain information, are motivated to do so and believe they are fulfilling their duty, while they are disapplying the law: unintentional disobedience. If people are unable to perform the exclusion task when aware of it and motivated to do so, the advantage of the exclusionary model over the weighing model is diminished. This possibility deserves attention.[56]

66 As far as the relevance element is concerned, we do not know how many of the judges who acquitted after having declared the testimony inadmissible considered that it would be right — in a world without exclusionary reasons — to admit the testimony. That is, whether for them the testimony was a reason or a quasi-reason. It may be psychologically interesting to know whether the source of legal contamination is the conscious or the unconscious morality of the judge. In any case, this uncertainty doesn't diminish the challenge for the exclusionary model.

67 The strength of this experiment lies in the fact that some participants who apparently failed the exclusion task were fully aware that they were involved in an exclusion task, although we cannot be sure how aware they were of the biasing character of the information. The behaviour of judges who admitted evidence may be interesting from the point of view of criminal law. But what is interesting from a psychological point of

view is the behaviour of those judges who excluded the evidence but behaved like those who admitted it. At least these judges believed that the law required them to exclude certain information. And provided that the awareness element is clearly present in the experiment, the possible lack of the relevance of the information does not weaken the critique of the exclusionary model so much as it strengthens it: if they are not able to exclude irrelevant elements *a fortiori* they will not be able to exclude what for them are reasons for action. Moreover, it is legitimate to imagine that even for those who consider testimony irrelevant already on the moral level, testimony appears as something of practical relevance, certainly more so than random numbers in anchoring (a quasi-reason). This makes it easier to assume that many judges not only considered the information inadmissible, but some also had an awareness of its biasing character.

68 What is missing is clear proof of the motivation to follow the exclusionary model rather than the weighing model or the simple balancing of background reasons. The fact that judges recognized that they are faced with an exclusion task and that they declared to have fulfilled the task, while we know they didn't, is consistent both with the thesis that they have been unconsciously conditioned by the information, and with the thesis that they intentionally opted for balancing the reasons. In one case we are in front of a failure to follow the exclusionary model, in the other, it is the case that judges opted for weighing (what they believed to be) the background reasons, or for the weighing model properly called (background reasons + authority of law as first order reasons). This possibility is open, but taking it seriously is tantamount to saying that the experiment is completely unable to demonstrate anything particularly psychologically relevant.[57] If we consider the experiment relevant — until someone proves the contrary — then we must consider it as a proof of a "failure to deliberately disregard relevant information", that is a failure of the exclusionary model.

69 The previous discussion illustrates how experimental results can undermine confidence in the exclusionary model to counteract bias. In support of this conclusion, it is also interesting to analyse another experiment with a different structure.

70 According to the harm principle applied to criminal law, only harmful conduct deserves to be sanctioned through criminal punishment. The hypothesis studied by the two authors of the experiment in question, Avani Metha Sood and John Darley, is that when people want to punish a certain behaviour that is not harmful and they are presented with a legal constraint reflecting the harm principle, they tend to perceive the behaviour as harmful, expanding their conception of harm — the plasticity of harm hypothesis. Again, the manipulation is not conscious. Rather, people genuinely believe that their judgements on concrete cases are an application of the harm principle, when instead their perception of harm is inadvertently influenced by their goals — the illusion of objectivity.[58] The experiment assumes, in other words, the occurrence of the previously mentioned phenomenon of motivated reasoning.

71 In the first phase of the experiment, a behaviour is identified that many people would like to criminalise even though they believe that this behaviour is harmless.[59] One of the conducts selected by the authors in this first phase of the experiment is "going naked to the supermarket". Going to the supermarket naked is thus a conduct that leads people to infringe the harm principle because according to the harm principle only harmful conduct may be criminalized.

72 The second part of the experiment is aimed at stimulating a process of motivated reasoning and proving the "plasticity of harm hypothesis".[60] People are divided into

two groups, the experimental group and the control group. Both groups are presented with the story of the person going to the supermarket naked, but the experimental group is offered additional legally relevant information. They are told that "U.S. courts have decided that the government can impose a criminal penalty only upon conduct that is shown to cause harm".[61] In essence, it is stated that the courts have ruled that criminal law must always respect the harm principle: "the necessity-of-harm constraint". This is not necessarily realistic, but the authors of the study make sure that the people tested believe that this is a true principle established by U.S. courts. In addition, both groups are given a definition of "harm", whereby "Harm, for these purposes, is defined as injury to a person or persons that can be clearly demonstrated. There could be types of conduct that are wrong, but do not cause harm".[62] Both groups are then asked the following questions: (i) whether the government should repress the conduct under consideration through criminal law; (ii) whether the conduct caused a demonstrable harm; (iii) how harmful the conduct was.

73 Well, if law had a causal impact on the decision, then (i) the rate of perceived harmfulness of the conduct should remain unchanged between the control and experimental group; and consequently (ii) the rate of criminalization of the conduct in the control group should plummet in the experimental group. Instead, the result of the experiment is that in the experimental group, the rate of criminalization stays constant even though people are subject to the necessity-of-harm constraint, and the number of harm reports more than doubled in the constraint condition as compared to the control condition.[63] By being subjected to an argumentative constraint, people perceive conduct as harmful and criminalize it on the basis of this perception. This should prove the hypothesis according to which people, without being aware of it, are able to manipulate their conception of harm to achieve the desired result.[64] We can say that law has a causal efficacy on people, but this efficacy does not concern the decision, which remains unchanged, as much as the reasoning that leads to the decision; reasoning that does not reflect the true reasons leading to the conviction and is therefore a form of *ex post* rationalisation.

74 In this case, according to the Razian framework, the ruling of U.S. courts is a first-order reason not to criminalise harmless conduct and a second-order reason excluding reasons for and against criminalization. The structure of the experiment ensures that people are aware of the exclusion task they are facing. People's attention is drawn to the exclusionary task twice: first when they are told about the ruling of U.S. courts and second when they are given a conceptual scheme according to which there are wrongs that are not harms. Moreover, in this case it is quite plausible that the element of relevance is also present. It is plausible that not only does going to the supermarket naked trigger one or many of the subjects' moral foundations, but also their conscious moral values, as was demonstrated through the first phase of the experiment in which the authors found conduct that for many deserves criminal punishment even if it is not harmful. Finally, it must be remarked that in this experiment the authors also attempted to prove the element of motivation, by proving that the subject responded "under the illusion of objectivity".[65]

75 On these grounds, the experiment questions the advantages of the exclusionary model over the weighing model. Thanks to the responses recorded in the first phase of the experiment, we know that for many people the result of the balancing of the underlying reasons is that the naked person should be punished through criminal law.

In contrast the responses of the people in the constraint condition should represent the application of the exclusionary model. According to this model, the naked person's behaviour should not be criminally punished,[66] yet, on the face of it, it appears that well-motivated persons, aware of the task of exclusion, are unable to neutralise their moral deliberation.

7 Doubts and conclusions

76 *If* the experiments seen in the previous section prove the experimenters' hypothesis, then we have reason to believe that there is such a thing as involuntary disobedience to law. As Sood nicely puts it, elaborating on a position expressed by Robert Cover, a judge who is caught between law and morality seems to have only four options: to apply the law against morality; to follow morality rather than the law; to resign; and to cheat, that is to resolve the conflict between law and morality by resorting to an interpretation of the law she does not believe. To these options, Sood notes, we must add a fifth in which "judges may unintentionally construe facts and apply law in a way that preserves the appearance— not only to others but also to themselves – of conforming to both the legal doctrine and their own intuitions about justice".[67]

77 I argued that the exclusionary model may be challenged by the phenomenon of unintentional disobedience, pointing to a general psychological explanation, source confusion, and to examples of experiments that can be used to assess the magnitude of this phenomenon. Again, the fact that the exclusionary model suffers from its own problems related to people's inability to complete an exclusionary task does not mean that the weighing model is preferable. Even if limited to the goal of limiting bias, the exclusionary model might still outperform the weighing model. However, we may have reason to believe that the exclusionary model is far from satisfactory, and that we should prefer a different model, e.g., something along the lines of Gur's dispositional model, according to which law meeting certain requirements gives agents reasons to cultivate a law-abiding disposition; a model that does not rely on mental performance of dubious effectiveness. Once in focus, this conclusion must be subjected to critical examination.

78 First of all, it must be said that Wistrich, Guthrie, and Rachlisnki themselves, in the study reported above, note how in several cases — especially those involving rights protected by the U.S. Constitution — judges effectively neutralise the impact of inadmissible information on their reasoning.[68] The extent of the failure of the exclusion task is therefore still to be explored.

79 Second, it should be noted — for the moment only *en passant* — that the category of involuntary disobedience resolutely clashes with a certain way of understanding the law. On some views, legal provisions are indeterminate: "Judging requires applying loosely defined concepts or broad standards to a concrete case, tailored to its circumstances through interpretation".[69] Raz's exclusionary model often seems tailored to cases that do not raise great interpretative problems. If we take into account legal indeterminacy and interpretative problems, it is no longer obvious on what basis it is possible to censure the application of the rape-shield statute or the manipulation of the harm principle as cases of involuntary disobedience to law. Consider, in the experiment about the harm principle, the thesis according to which when people criminalize the conduct of going naked to the supermarket they are manipulating the concept of harm.

Against this idea someone may retort by observing that reasonable people may interpret this conduct as a source of harms, because, for example, "children may see it and this will hurt them", or "this is offensive behaviour, and offensive behaviour is harmful" or "such behaviour may stir people's sexual desires in a way that could cause harm". This does not mean that the authors of these studies have neglected the entire issue outright. In the experiment about the manipulation of the harm principle the first part of the experiment is concerned about the individuation of a conduct that people *spontaneously* take to be harmless. This part aims precisely to counter the objection according to which people simply disagree about what conduct is harmful. Thanks to this part of the experiment, it is plausible to believe that people formulated an exclusion task and then failed to perform it. Still, it remains possible that going to the supermarket naked may be sensibly interpreted as harmful conduct because children may be hurt, because offensive behaviour may be harmful, and so on. How exactly the experimental results in favour of unintentional disobedience can be reconciled with the idea of legal indeterminacy and the problem of interpretation is something that requires further elaboration.[70]

80 However, even with these limitations, this paper develops the integration process between Raz's theory of authority and cognitive science studies. In this regard, it should be mentioned that Gur's analysis of biases is not limited to the weighing model. He believes that the exclusionary model also fails to counteract biases to a successful degree. For my analysis to contribute to the debate I must explain where the problem of mental contamination differs from Gur's critique of the exclusionary model. Gur general idea is that the exclusionary model "better enables law to counteract those biases than the weighing model does, since the latter, unlike the former, invariably requires subjects to act through all- things- considered balancing exercises".[71] The reason why Gur believes that the exclusionary model doesn't counteract biases to a successful degree revolves around the *piecemeal* structure of Raz's theory of authority, which is derived by the so called "normal justification thesis", according to which

> the normal way to establish that a person has authority over another person involves showing that the alleged subject is likely better to comply with reasons which apply to him (other than the alleged authoritative directives) if he accepts the directives of the alleged authority as authoritatively binding and tries to follow them, rather than by trying to follow the reasons which apply to him directly. (Raz, 1986: 53)

81 This means that to test whether one person has authority over another we need to test the advantage of obedience on an individual basis and in a domain-specific manner. Following this idea, for example, the expert pharmacologist is not subject to the authority of laws about the safety of drugs. But, as Gur notes, if the scope of legitimate authorities is so patchy, people willing to follow the exclusionary model are asked to act "under unfavourable conditions that involve potential exposures to biases".[72] First, their assessment of superiority over the person claiming authority is subject to bias, as is their balancing of reasons. Second, even in the cases in which the subject is really superior as a matter of expertise than the person claiming authority, she will be affected by other biases not linked to epistemic superiority — biases that have "little to do with the lack of information".[73]

82 Gur's critique revolves around a peculiarity of Raz's theory of authority — the piecemeal character of authority — whereas the critique offered in this article challenges the exclusionary model in general. While the problem Gur stresses is that

there will be too many occasions in which people will decline to accept that they have reasons to perform an exclusion task, mine is that there will be too many occasions in which people who are motivated to perform the exclusion task will fail to do so successfully.

83 Even more generally, we may distinguish three cases in which biases lead to wrongful disobedience in the context of the exclusionary model.

> (1) Facing an unsettling order, the agent is driven — contrary to reasons — to believe that the superior lacks authority.
>
> (2) Facing an unsettling order, the agent is driven — contrary to reasons — to believe that there are reasons against the order that are outside the scope of exclusion of the authoritative directive (this is the case Gur analysed).

84 Besides these two cases of disobedience contrary to reasons, we may add a third case featuring mental contamination:

> (3) Facing an unsettling order, the agent is driven — contrary to reasons — to interpret the order in a non-unsettling way.

85 The difference between this case and the other two is that, while in the others the agent is aware she is not obeying the superior (though she believes she is responding to reasons), in this last case the agent mistakenly believes she is obeying to the superior. This is the case of involuntary disobedience, the case in which the subject believes that she must compartmentalize her thoughts — severing morality form legality, in the sense specified above — and she is motivated to do so, but the compartmentalization fails.

86 Finally, I would like to draw attention to an answer or, perhaps better, a possibility of metamorphosis that remains open for the exclusionary model. As mentioned above, Raz's said that "possibly it *[reflection on the merits of actions]* could be prohibited by a special directive to that effect."[74] Studies on mental contamination show that we are sometimes unable to counter biases despite our awareness and motivation. If so, our mental power to exclude (what we regard as) first-order reasons defeated by exclusionary reasons is fallible. So, how can we counter biases more efficiently? There are several possibilities, like the cultivation of the disposition to obey the law (this is the core idea of Gur's dispositional model[75]), or investments in choice architecture[76] leading to a real compartmentalization of information (measures that are already common in the adversarial system in criminal law). To provisionally summarize all these measures under one label, we may say that a possible solution to the problem of mental contamination is to exploit *pre-commitment strategies*, which in Elster's terms may be defined as any decision adopted at time T1 that increases the probability that one will carry out another decision at time T2.[77]

87 That said, we can conceive of a version of the exclusionary model that is not reduced to the exclusion task through mind control, but that incorporates pre-commitment strategies. An open question that deserves further investigation is the extent of this metamorphosis, which may not only affect the decision-making strategy required by the model but go deep into its very nature.

88 A distinction can be made between models about practical reasons and models about practical reasoning.[78] The former are concerned with answering questions about what reasons for action exist and how they interact; the latter are concerned with prescribing decision-making processes. As far as we interpret the exclusionary model as prescribing a specific mental performance — the exclusion task — we see it as a model about practical reasoning. But it would cease to be a model of practical reasoning and become a theory of practical reasons if we took the model as the assertion that there are sometimes reasons not to act on certain reasons, without any specific commitment to a specific decision-making process to arrive at the correct practical solution. A theory which, in other words, indicates just the goal of the exclusion of reasons, legitimising a wide variety of strategies that could lead to this result.

89 This is not the context in which it is possible to address the issue of which understanding of the exclusionary model is better, or in general, the respective advantages of each interpretation of it. It must be said as a conclusion that this is not merely a philological question about Raz's thought, which already be of interest on its own anyway. Precisely because Raz laid the foundations for a wide-ranging debate on the authority of law, it should come as no surprise that the concepts he introduced could be used to elaborate problems other than those strictly addressed in the corpus of his work.

BIBLIOGRAPHY

Bobbio, N. (1984). Governo degli uomini o governo delle leggi? In N. Bobbio, *Il futuro della democrazia* (pp. 157-180). Einaudi.

Bublitz, J. C. (2020). What is wrong with hungry judges? A case study of legal implications of cognitive science. In A. Waltermann, D. Roef, J. Hage & M. Jelcic (Eds.), *Law, Science and Rationality* (pp. 1-30). Eleven.

Celano, B. (2003). Are reasons for action beliefs? In L.H. Meyer, S.L. Paulson & T.W. Pogge (Eds.), *Rights, Culture and the Law. Themes from the Legal and Political Philosophy of Jospeh Raz* (pp.25-43). Oxford Univeristy Press.

Celano, B. (2021). *Lezioni di filosofia del diritto. Costituzionalismo, stato di diritto, codificazione, diritto naturale positivismo giuridico. Seconda edizione ampliata*. Giappichelli.

Celano, B. (2023). Dog Law. The logical structure of distinguishing (or its lack thereof). In T. Endicott, H. D. Kristjánsson & S. Lewis (Eds.), Philosophical Foundations of Precedent (pp. 214-227). Oxford University Press.

Chang, R. (2016). Comparativism: The Grounds of Rational Choice. In E. Lord & B. Maguire (Eds.), Weighing Reasons (pp. 213-240). Oxford University Press.

Damaška, M. R. (2003). Epistemology and Legal Regulation of Proof. *Law, Probability and Risk*, *2* (2), pp.117-130.

Danzinger, S., Levav J. & Avnaim-Pesso L. (2011). Extraneous Factors in Judicial Decisions. *Proceedings of the National Academy of Science, 108*(17), pp. 6889-6892.

Ditto, P. H., Pizarro, D. A., & Tannenbaum, D. (2009). Motivated moral reasoning. In D. M. Bartels, C. W. Bauman, L. J. Skitka, & D. L. Medin (Eds.), Moral judgment and decision making (pp. 307–338). Elsevier Academic Press.

Elster J. 1984 [1979]. *Ulysses and the Sirens. Studies in Rationality and Irrationality. Revised Edition*. Cambridge University Press.

Fuller, L. L. (1964). *The Morality of Law*. Yale University Press.

Gilbert, D. T. (2002). Inferential correction. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), Heuristics and biases: The psychology of intuitive judgment (pp. 167–184). Cambridge University Press.

Gilovich, T., Griffin, D., & Kahneman, D. (Eds.). (2002). Heuristics and biases: The psychology of intuitive judgment. Cambridge University Press.

Glöckner, A. (2016). The irrational hungry judge effect revisited: Simulations reveal that the magnitude of the effect is overestimated. *Judgment and Decision Making*, *11*(6), 601-610.

Golding, J. M., & Hauselt, J. (1994). When Instructions to Forget Become Instructions to Remember. *Personality and Social Psychology Bulletin*, *20*(2), 178-183.

Gur, N. (2018). *Legal Directives and Practical Reasons*. Oxford University Press.

Haidt, J. (2001). The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment. *Psychological Review, 108*(4), 814-834.

Haidt, J. (2012). *The Righteous Mind. Why Good People Are Divided by Politics and Religion.* Penguin Books.

Haidt, J. & Hersh, M.A. (2001). Sexual Morality: the Culture and Emotions of Conservatives and Liberal. *Journal of Applied Social Psychology, 31*(1), 191-221.

Haidt, J. & Bjorklund, F. (2008). Social Intuitionists Answer Six Questions about Moral Psychology, In W. Sinnott-Armstrong (Ed.), *Moral Psychology. Vol II* (pp. 181-217). MIT Press.

Hart, H.L.A. (1982). Commands and Authoritative Legal Reasons. In H.L.A. Hart, Essays on Bentham: Jurisprudence and Political Philosophy (pp. 243-268). Oxford University Press.

Hurd, H. M. (1991). Challenging Authority. Yale Law Journal, 100(6), 1611-1677.

Kahneman D., Tversky A. (1974). Judgment Under Uncertainty. Heuristics and Biases. *Science, New Series,185*(4157), 1124-1131.

Kunda, Z. (1990). The case for motivated reasoning. Psychological Bulletin, 108(3), 480–498.

Maroney, T.A. (2011). The Persistent Cultural Script of Judicial Dispassion. *California Law Review*, 99 (629), 629-682.

Moore, M. (1989). Authority, Law, and Razian Reasons. Southern California Law Review, 62, 827-896.

Pyszczynski, T, & Greenberg, J. (1987). Toward and integration of cognitive and motivational perspectives on social inference: A biased hypothesis-testing model. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 20, pp. 297-340). Academic Press.

Raz, J. (1986). *The Morality of Freedom*. Oxford University Press.

Raz, J. (1989). Facing Up: A Reply. *Southern California Law Review, 62*(3 & 4), 1153-1236.

Raz, J. (1990). *Practical Reasons and Norms*. Oxford University Press (ed. or. 1975).

Raz, J. (2009a). Legitimate Authority. In J. Raz, *The Authority of Law* (second edition, pp. 3-27). Oxford University Press (ed. or. 1978).

Raz, J. (2009b). The Rule of Law and its Virtues. In J. Raz, *The Authority of Law* (second edition, pp. 210-229). Oxford University Press (ed. or. 1977).

Rocché, G. & Ubertone M. (forthcoming). Can Disgust Predict Legal Decision-Making? An Experimental Jurisprudence Perspective on Gut Feelings and the Rule of Law. *Isonomía. Revista de teoría y filosofía del derecho.*

Ross, L., Lepper M.L. & Hubbard, M. (1975). Perseverance in Self- Perception and Social Perception: Biased Attributional Processes in the Debriefing Paradigm. *Journal of Personality and Social Psychology, 32*(5), 880-892.

Schauer, F. (2004). The Limited Domain of the Law. Virginia Law Review, *90*(7), 1909-1956.

Shapiro, S. (2002). Authority. In J. C. Coleman & S. Shapiro (Eds.), *The Oxford Handbook of Jurisprudence and Philosophy of Law* (pp. 382-439). Oxford University Press.

Sood, A.M., & Darley, J.M. (2012). The plasticity of harm in the service of criminalization goals. *California Law Review*, *100*, 1313-1358.

Sood, A.M. (2015). Cognitive Cleansing: Experimental Psychology and the Exclusionary Rule. *Georgetown Law Journal*, *103*, 1543-1608.

Strohmeier, N. & De Jong, S. (2023). Moral Character Judgments and Motivated Cognition in Legal Reasoning. *Diritto & Questioni Pubbliche/RECOGNISE Legal Reasoning And Cognitive Science: Topics And Perspectives, Special Publication / August*, 227-248.

Thaler, R. H., & Sunstein, C. R. (2008). *Nudge. Improving Decisions about Health, Wealth, and Happiness*. Yale University Press.

Vassiliou, A. (2022). The Normativity of Law: Has the Dispositional Model Solved our Problem? Oxford Journal of Legal Studies, *42*(3), 943–962.

Wilson, T.D., Brekke N. (1994). Mental contamination and mental correction: Unwanted influences on judgments and evaluations. Psychological Bulletin, 116(1), 117–142.

Wilson, T.D., Centerbar, C.D., & Brekke, N. (2002). Mental Contamination and the Debiasing Problem. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), Heuristics and biases: The psychology of intuitive judgment (pp. 185-200). Cambridge University Press.

Wistrich, A.J., Guthrie C. & Rachlinski J.J. (2005). Can Judges Ignore Inadmissible Information - The Difficulty of Deliberately Disregarding. *University of Pennsylvania Law Review, 153*(4), 1251 – 1345.

Wistrich, A.J. & Rachlinski, J.J. (2017). Implicit Bias in Judicial Decision Making How It Affects Judgment and What Judges Can Do About It. *Chapter 5: American Bar Association, Enhancing Justice (2017), Cornell Legal Studies Research Paper No. 17-16*, 87-130.

NOTES

1. Raz 1999 [1975]: 36.
2. Raz 1999 [1975]: 39.

3. Raz 1999 [1975]: 191; for the expression "pre-emptive reasons" see Raz 1985: 42, 46; cfr. Gur 2018: 13, and Moore 1989: 853.
4. Gur 2018: 16.
5. Raz 1986: 53.
6. Raz 1986: 61-62.
7. On the assumption that the actual location of the threat is one of the reasons inside the scope of the exclusion. This may be disputed, but these disagreements are immaterial for the present discussion.
8. Raz 1999 [1975]: 38.
9. This way of referring to Raz's theory of authority has been introduced by Vassiliou 2022: 843.
10. Gur 2018: 15 and the literature cited; see also Hurd 1991: 1641 ff.; and Schauer's rule-sensitive particularism (Schauer 1991: 97).
11. This idea in the Italian context has been illustrated by Bobbio 1984: 157 ff.; and more recently by Celano 2021: 109-116.
12. I am not suggesting that the struggle against biases was central in Raz's theory. Raz argued that the basic argument in favour of authority lies in the need to overcome epistemic deficiencies due to lack of expertise and to solve coordination problems (Raz 1999 [1975]: 195; Gur 2018: 101). It is also true that Raz is not unaware of the problem of bias, since he mentions it among the five most common reasons to establish the authority of states (Raz 1986: 75; for other comments on the problem of biases see also Raz 1989: 1192). In general, I believe that even though Raz's reflections did not focus on the problem of biases — probably for historical reasons — but on other human limitations, today the problem of biases cannot but be considered as an important foundation of the exclusionary model, without betraying the spirit of Raz's work.
13. Gur has also analysed the problem of biases in relation to the exclusionary reason model (see Gur 2018: 127-130). We will return to this point later.
14. According to this analysis the exclusionary model is a normative model about the type of reasons we have. Doubts may be raised about this understanding of Raz's theory. There are passages in which Raz clearly states that he is interested primarily in the conceptual analysis of reasons rather than on the development of a substantive account (see Raz 1999 [1975]: 10). Still, Raz is also defending the idea that someone is a legitimate authority when she is able to direct people's behaviour to conform better with reasons (Raz 1986: 53 ff.), that is, he is defending the idea of authority against the challenge of philosophical anarchism, a form of error-theory of the duties to obey (see Shapiro 2002: 384-393, 408). In this context he argues that only if we adopt the exclusionary model the authority will be able to perform its service, and this means that there are circumstances in which exclusionary reasons exist (they are valid) (Raz 1999 [1975]: 195), a claim that cannot be established by only saying that we have the concept of exclusionary reason (see Gur 2018: 98). I find no contradiction in this project. On one hand, Raz believes that only if authoritative directives are exclusionary reasons, will obedient people succeed in conforming better to reasons (Raz 1986: 61). So, authoritative directives *ought* to be treated as exclusionary reasons. On the other hand, treating authoritative directives as exclusionary reasons reflects – for Raz – our common way to see authority, the phenomenology of authority. In this sense Raz's view is a form of conservatism: we have certain practices, certain concepts, certain experiences regarding authority, and it can be demonstrated that at least in principle these practices, concepts, and experiences are not irrational (it may happen that someone has legitimate authority giving us exclusionary reasons). Note that Raz explicitly states that the service conception is a normative model and dispels doubt about the confusion between conceptual analysis and normative arguments that supposedly plague his system (Raz 1986: 63). In his book Gur (2018) draws attention to both aspects of Raz's theory. Raz's conceptual aim is discussed in chapter V and turned into a phenomenological argument against the weighing model; while

Raz's normative stance, according to which we ought to treat authoritative directives as exclusionary reasons to make them fulfil their function, is discussed in chapter VI.

15. See, also, Raz 2009a [1978]: 25-27. The position according to which obedience doesn't imply surrender of judgment is contrasted by Raz with Hart's view (Hart 1982: 253).

16. I borrow this expression from Rachlinski & Wistrich 2017: 94.

17. A clarification might be helpful to avoid some misunderstandings. According to Raz's preferred approach, reasons are facts and not mental states (for a critical analysis, see Celano 2003). The failure of the exclusion task is the failure to neutralise mental states. There would seem to be an inconsistency between Raz's approach and the one proposed here. However, it is clear — first for Raz — that for reasons for action to guide conduct, they must be embedded in some mental state (Raz 1975: 17). We can only exclude reasons that are facts if we are able to exclude our mental states — appropriate or not — about these facts. Thus, the inability to neutralise the influence of certain mental states implies the failure of the exclusionary model.

18. Danzinger, Levav & Avnaim-Pesso 2011.

19. In general, see Kahneman & Tversky 1974.

20. Wistrich & Rachlinski 2017: 94.

21. I.e. our ability to exclude what we take to be a reason (see footnote 17).

22. Again, for "information that is reason" I mean "information that the agent takes to be a reason", since even though reasons are facts, only beliefs and proattitudes about reasons for action are factors in the decision-making process.

23. Therefore, the idea of quasi-reasons should not be confused with *prima facie* reasons.

24. The above applies in relation to the agent's point of view, which is what is of interest here. From the external point of view, the distinction between quasi-reasons and non-reasons may be useful to isolate cases where it is reasonable to be in doubt as to whether something was a reason for the agent or not, from cases where the doubt is not legitimate. It is indeed unlikely for an agent to think that random numbers are a reason to impose one punishment rather than another, but not so unlikely for him to think that his boss's aggressiveness is a reason to react.

25. See the title of Part B – "Anchoring, Contamination, and Compatibility" – of Gilovich, Griffin & Kahneman 2002, in particular Gilbert 2002 and Wilson, Brekke & Centerbar 2002.

26. Wilson & Brekke 1994: 128-130, 131-133. I will not try to establish here whether the problem for the exclusion task is more related to a failure to detect the direction and magnitude of the bias or to have mental control over our responses.

27. See Gilbert 2002: 180; Wistrich, Guthrie & Rachlinski 2005: 1264.

28. Damaška 2003.

29. Golding & Hauselt 1994.

30. See generally Ross, Lepper & Hubbard 1975.

31. It is possible that in Raz's theory it would be more appropriate to say that reason has been *cancelled* rather than excluded. I do not attempt to develop this idea here, because in any case there is no reason to imagine that we have two distinct mental powers, one to exclude and one to cancel. So, for the purposes of this paper, I will ignore this distinction. For the notion of cancelling reason see Raz 1999 [1975]: 27.

32. The expression "layered morality" is not taken from Haidt's theory, but I am convinced that it is useful in conveying the image of morality having a conscious and an unconscious component.

33. Haidt & Bjorklund 2008: 203. Haidt will distinguish later the fairness foundation and the liberty/oppression foundation (see Haidt 2012: ch. VIII).

34. Haidt 2012: ch. VII.

35. See Haidt 2001, and Haidt 2012: ch. II.

36. This aspect is discussed at length in Rocché & Ubertone (forthcoming): par. 2.

37. See in particular Haidt & Hersh 2001.

38. "Why *Good* People Are Divided by Politics and Religion" (italics added): This is the telling subtitle of Haidt's most famous book.

39. Kunda 1990: 480, 482 ff. According to Kunda's framework the phenomenon we are considering is the "directional version" of motivated reasoning. See also Strohmeier & De Jong 2023:228.

40. Ditto, Pizarro &Tannenbaum 2009.

41. Kunda 1990: 483; see for the introduction of the expression Pyszczynski & Greenberg 1987.

42. Sood & Darley 2012: 1322; but the closeness between the two models is emphasized by Ditto, Pizarro & Tannenbaum 2009: 313 ("From an intuitionist perspective, moral reasoning is, fundamentally, motivated reasoning").

43. "Moral emotions are one type of moral intuition, but most moral intuition are more subtle; they don't rise to the level of emotions (...) *Intuition* is the best word to describe the dozens or hundreds of rapid, effortless moral judgments and decisions that we all make every day. Only few of these intuitions come to us embedded in full-blown emotions" (Haidt 2012: 53).

44. Our personal assessment of background reasons: solitary moral reflection; or moral reflection in a social context, which is not institutionalized.

45. See for example Schauer 2004; and Maroney 2011.

46. Fuller 1964: 81; Celano 2023: 214.

47. Raz 2009b [1977]: 213.

48. See Moore 1989: 839 ff. As Moore notes, this continuity is one of the main differences between Raz's philosophical programme and Hart's effort to claim a space for legal obligation, making it independent of moral theory.

49. Respect for authorities is also one of Haidt's six moral foundations, and in this sense it is a component of our unconscious morality, the first layer of morality. So why am I developing a third layer of morality based on authority? Sensibly, an anonymous referee pointed out this tension in my analysis. The reason is that authority is paradoxical, because, I argue, it sometimes triggers our intuitions in favour of certain actions and other times makes us do what we consider deeply counter-intuitive. An agent may find it intuitively appropriate to bend the knee in front of a person he recognizes to have authority. And the same agent may believe that she must obey repulsively counterintuitive orders from an authority. Both behaviours have to do with authority. The problem of the relationship between authority, intuitions, and counter-intuitiveness is intriguing and deserves a systematic analysis that cannot be done here. Still, two preliminarily points must be emphasized. First, Haidt's moral foundations theory is related to the foundations of people's *intuitive* morality (Haidt & Bjorklund 2008: 203), and it is well suited to explain our intuitive judgments triggered by authority, but not the counter-intuitive judgments imposed by it. The agent who believes that she must bite the bullet because "orders are orders and should be obeyed even if wrong" is instead crucial in Raz's theory (Raz 1999 [1975]: 38). And in fact — second issue — Raz's service conception is not built on an innate appreciation of hierarchies, and can well be grounded just on the most liberal-progressive foundations like the care/harm foundation and the liberty/oppression foundation. So, at first glance there are two types of recognition of authority at play here, one intuitive — the Haidtian — the other cognitively costly — the Razian.

50. Wistrich, Guthrie & Rachlinski 2005: 1298 ff.; see also Wistrich & Rachlinski 2017: 95.

51. Information about the sexual conduct is in my terminology an example of "quasi-reasons". In this case mere fact that there is a debate about their relevance and the direction of their relevance shows that their perception is very different from that of random numbers.

52. Wistrich & Rachlinski 2017: 95.

53. Wistrich, Guthrie & Rachlinski 2005: 1301.

54. Wistrich & Rachlinski 2017: 95.

55. As we said in the first section, the weighing model does not entirely deprive authoritative directives of normative relevance, since it treats them as content-independent first-order reasons.

56. The rule on inadmissibility of evidence has a special feature in that the operation prescribed by the rule consists in depriving certain facts of normative relevance. In other words, we are not faced with a rule that is a first-order reason to do x and a second-order reason not to consider reasons not to do x (or to do x). A possible reconstruction of the structure of this norm in Razian terms is that it produces two exclusionary reasons. The first exclusionary reason is the reason for not weighing the pros and cons when it comes to deciding on the admission of a certain piece of evidence. The second exclusionary reason is the reason for not admitting certain evidence in the trial, a reason for not finding a certain person guilty or innocent for certain reasons. I have to thank Michele Ubertone for a clarifying discussion on this point. Another issue related to the peculiarity of this rule is that it may be a source of reasons for belief rather than of reasons for actions. I am indebted, this time, to Yahya Gülgeç for this observation. The discussion of the structure of such rules from the perspective of Raz's theory deserves attention, and it should be the subject of a separate work focusing on the different types of legal contamination. Suffice it to say that the particularity of this rule does not seem to invalidate the relevance of the experiment for the purpose of proving our inability to perform an exclusion task.

57. As an anonymous reviewer noted, caution is called for when referring to empirical data, since even important studies, like Danzinger's study on hungry judges, have been questioned in terms of their empirical relevance (Glöckner 2016). Later, in addressing Bublitz's research, I will dwell on another possible challenge, which is instead fully theoretical.

58. Sood & Darley 2012: 1321.

59. Sood & Darley 2012: 1325 ff.

60. Sood & Darley 2012: 1328 ff.

61. Sood & Darley 2012: 1328.

62. Sood & Darley 2012: 1328.

63. Sood & Darley 2012: 1333.

64. Sood &Darley 2012: 1324 f.

65. See Sood & Darley 2012: 1342 ff.

66. "Why not?" someone might ask. Given the indeterminacy of legal concepts, a reasonable person may interpret the conduct harmful. But we must remember that the experimenters did not overlook this worry. This is why the authors selected through a preliminary empirical study, conduct that is not normally interpreted as harmful.

67. Sood 2015: 1563 f.

68. Wistrich, Guthrie & Rachlinski 2005: 1259, 1322.

69. Bublitz 2020: 15 f.

70. According to Bublitz, legal errors must be divided between *explicit* errors, which can be found in the legal opinions laid down by the judges, and *imperceptible* errors, which "cannot be extracted from records because they are not contained in them" (Bublitz 2020: 6 ff.). In this framework, some biases are a form of imperceptible errors, and a source of imperceptible errors is, of course, legal indeterminacy. Indeterminacy, then, makes it possible that a decision that is

not explicitly flawed is in some sense wrong (Bublitz 2020: 15 f.), but it is unclear where this wrongness stems from — an issue that Bublitz tries to tackle. I am particularly thankful to an anonymous referee for having raised the problem of indeterminacy and pointing out Bublitz's recent paper. The examples in quotations marks are the same used by the referee to criticize the experiment about the manipulation of the harm principle. The referee believes that Bublitz' account might challenge my analysis. Bublitz's work is focused on Danzinger's famous study on hungry judges (Danzinger, Levav & Avnaim-Pesso 2011). It must be noted that Bublitz is not casting doubt on the reliability of empirical findings as other authors did, rather he is criticising the normative implications drawn from the experiment. In particular Bublitz argues that framing the impact of blood glucose level and mental fatigue as a problem of "'extralegal factors' contaminating decision is misleading" (Bublitz 2020: 6). Bublitz is contrasting extralegal factors meant as extralegal reasons "considerations which courts should not take into account" (Bublitz 2020: 10) with glucose level and mental fatigue, which are simply causes or technical limitations (Bublitz 2020: 10 ff. 22). A parsimonious reply consists in noting that the idea of moral contamination and the experiments I have chosen are relevantly different from Danzinger's study because they regard features of the conduct to be judged and not technical limitations of human decision-making. Therefore, they are good candidates for extralegal factors contaminating decision. On the other hand, Bublitz's critique may be widened so to embrace also these influences. At times even Bublitz seems to do this, like when he understands some racial biases as violations of the principle of equal treatment — a problem he contrasts with the idea of the extralegal factors. Be that as it may, it is not possible here to fully address Bublitz's position.

71. Gur 2018: 129 f.

72. Gur 2018: 128.

73. About this second problem Gur recognizes that Raz may retort by saying that in assessing the superiority of the person over the authority, we must take into account the problem of biases. Still, if we take into account biases to this extent then "the typical scope of legitimate authority no longer appears to be piecemeal in character and is therefore quite different from how Raz pictures it" (Gur 2018: 129).

74. Raz 1986: 39.

75. Gur 2018: ch. VII.

76. Thaler & Sunstein 2008.

77. Elster 1984 [1979]: 49; the concept of pre-commitment has already been explored in its connection with the theory of authority in Shapiro 2002: 418 ff. An analysis of Shapiro's position is beyond the scope of this paper.

78. Vassiliou 2022: 945 ff.; Chang 2016.

ABSTRACTS

According to Joseph Raz, authoritative directives are exclusionary reasons, which means that people have normative reasons not to act on their personal assessment of the merit of the case even though this doesn't imply that they have to surrender their personal judgment. The exclusionary model of authority is contrasted with the weighing model. One reason to accept the exclusionary model of authority over the weighing model is that human beings are fallible and obedience to the law can be a way to protect them from bias. This article questions the ability of the exclusionary model to counteract biases. To do so, it combines three different traditions of

empirical studies: a general framework on our inability to intentionally ignore relevant or quasi-relevant information — studies on 'mental contamination'; Jonathan Haidt's research on the conflict between moral reasoning and moral intuitions; and studies on motivated reasoning. While the overall picture does not doom the exclusionary model, it provides reason to articulate it as implying that surrender of judgment is often important.

Izločitveni razlogi in mentalna kontaminacija: *izziv za Razovo teorijo avtoritete*. Po Josephu Razu so avtoritativne direktive izločitveni razlogi, kar pomeni, da imajo ljudje normativne razloge, da ne ravnajo po svoji osebni vsebinski presoji okoliščin, čeprav to ne pomeni, da se ji morajo odpovedati. Model avtoritete, ki temelji na izločitvenih razlogih, je v nasprotju z modelom avtoritete, ki je utemeljen na tehtanju. Prvi model model naj bi imel prednost pred drugim drugim zato, ker smo ljudje zmotljivi, poslušnost zakonom pa naj bi nas ščitila pred pristranskostjo. V tej razpravi se to trditev postavlja pod vprašaj na podlagi treh različnih tradicij empiričnih raziskav. To so raziskave o "mentalni kontaminaciji", ki postavljajo splošni okvir o naši nezmožnosti namernega ignoriranja pomembnih ali kvazi pomembnih informacij; raziskave Jonathana Haidta o konfliktu med moralnim sklepanjem in moralnimi intuicijami; ter raziskave o motiviranem sklepanju.

INDEX

Keywords: Raz (Joseph), exclusionary reasons, surrender of judgment, biases, Haidt (Jonathan), moral foundations, motivated reasoning
motsclessl Raz (Joseph), izločitveni razlogi, odpoved presoji, pristranskost, Haidt (Jonathan), moralni temelji, motivirano sklepanje

AUTHOR

GIUSEPPE ROCCHÈ

Post-doc researcher, University of Palermo (Italy).
Address: Dipartimento di Giurisprudenza, Università degli studi di Palermo, Piazza Bologni 8, 90134 Palermo, Italy
E-mail: giuseppe.rocche@unipa.it.