Philos Stud (2022) 179:2187–2198
https://doi.org/10.1007/s11098-021-01759-1

Unjust combatants, special authority, and "transferred responsibility"

Luciano Venezia[1] · **Rodrigo Sánchez Brígido**[2]

Accepted: 29 October 2021 / Published online: 24 November 2021

Abstract Yitzhak Benbaji argues that those combatants who have agreed to blindly obey their superiors and who are ordered to fight in unjust wars are released from their duty to deliberate about the merits of the acts that they are ordered to perform. This is because their agreements result in the combatants' permissible lack of a necessary capacity for moral responsibility. Thus, the combatants are not morally responsible for their wrongful acts—their moral responsibility is "transferred" to their superiors. We argue, first, that Benbaji's own reasoning suggests that the agreements entered into between the combatants and their superiors are not binding and, second, that even if such agreements are binding, those combatants who obey their orders to fight are nevertheless morally responsible for their wrongful acts. Thus, Benbaji has failed to show that the combatants are permitted to act as ordered. By critically examining Benbaji's view, then, we defend the revisionist position that just and unjust combatants are morally unequal.

Keywords Authority · Contractual obligation · Jeff McMahan · Just war theory · Moral equality of combatants · Moral responsibility · Traditionalism · Tracing · Transferred responsibility · Revisionism · Yitzhak Benbaji

✉ Luciano Venezia
lucianovenezia@yahoo.com.ar; lvenezia@unq.edu.ar

[1] National University of Quilmes and National Scientific and Technical Research Council, Gurruchaga 160 2 I, 1414 Buenos Aires, Argentina

[2] National University of Cordoba and University of San Andrés, Gurruchaga 160 2 I, 1414 Buenos Aires, Argentina

1 Introduction

Suppose a set of combatants is ordered to fight in an unjust war. Many philosophers nowadays, notably those who support the revisionist view in just war theory, argue that these combatants are not permitted to obey their orders to fight.[1] This is because, according to this position, just and unjust combatants are morally unequal in the sense that, while just ones are permitted to kill unjust ones, unjust ones are not permitted to kill just ones.[2] The orders to fight, then, are not binding on those unjust combatants.

Although it is a prominent position, not all modern philosophers share the revisionist view. One important opponent is Yitzhak Benbaji, who specifically highlights the fact that combatants are members of hierarchical institutions and, also, that the correct functioning of these institutions requires a division of epistemic labor between those who form different links in the chain of command. In an important recent essay, he argues that those combatants who have agreed to blindly obey their superiors and who are ordered to fight in unjust wars are released from their duty to deliberate about the merits of the acts that they are ordered to perform. This is because their agreements result in the combatants' permissible lack of a necessary capacity for moral responsibility. Thus, they are not morally responsible for their wrongful acts—their moral responsibility is "transferred" to their superiors, so that "[a] military leader might be solely responsible for the [wrongful] killings that her subordinate commit[s] as a result of following her orders" (Benbaji, 2021, 3; see also 15–16). As non-morally responsible agents are, Benbaji submits, permitted to act as ordered, the combatants are then permitted to kill their opponents. Thus, Benbaji's account involves a defense of the traditional view that just and unjust combatants are moral equals, provided that the special cases that he considers are sufficiently widespread.[3]

Even if the cases that Benbaji focuses on are sufficiently prevalent in war (an empirical question that he leaves open and that we do not address), we argue, first, that his own reasoning suggests that the agreements entered into between the combatants and their superiors are not binding and, second, that even if such agreements are binding, those combatants who obey their orders to fight are nevertheless morally responsible for their wrongful acts. Thus, Benbaji has failed to show that the combatants are permitted to act as ordered. By critically examining Benbaji's view, then, we defend the revisionist position that just and unjust combatants are moral unequals.

Here is the plan. First, we introduce the analysis of moral responsibility adopted by Benbaji (Sect. 2). Next, we describe Benbaji's reasoning for the view that subordinates' moral responsibility is transferred to superiors in "special authority cases" (Sect. 3). After that, we put forward our criticisms. Benbaji analyzes this

[1] Unless explicitly indicated, we will use moral notions in the fact-relative sense.

[2] For defenses of the thesis that just and unjust combatants are not moral equals, see, most prominently, Fabre (2012), Frowe (2014), McMahan (2009), Rodin (2002).

[3] A related view, although much less developed, is put forward by Walzer (2004, ch. 2).

kind of case in light of two other kinds of case. We show that his reasoning suggests that the agreements entered into between combatants and their superiors in special authority cases are not binding. We also argue that, even if the agreements are binding, those subordinates who obey their orders are still morally responsible for their wrongful acts. At this point we also deal with a reply by Benbaji to our reasoning (Sect. 4). Section 5 brings the paper to a close.

2 Moral responsibility

Benbaji defends his account of transferred responsibility in special authority cases by relying on Jeff McMahan's criterion of moral responsibility for posing unjust threats of harm. This might come as a surprise, not only because McMahan is the most prominent revisionist just war theorist, but also because he explicitly endorses non-transferability.[4] Still, Benbaji submits that McMahan's position, if correctly modified in accordance with McMahan's own logic, implies that, in special authority cases, moral responsibility is transferred from subordinates to superiors.

McMahan's criterion is that a person is morally responsible for posing an unjust threat of harm if and only if she voluntarily chooses to engage in a risk-imposing activity and that activity will eventuate in harms to a victim who has a right not to suffer such harms.[5] To assign moral responsibility to a threatening person, then, it should be the case, first, that the threat that she poses results from her moral agency, which broadly speaking involves the capacity to exercise control over her own acts (McMahan, 2004, 724, cited in Benbaji, 2021, 4). In addition, moral responsibility for posing an unjust threat is sensitive to the degree to which the threat is foreseeable to the threatening person. Thus, if a person drives a car, "[she] will be [morally] responsible if, contrary to reasonable expectation and through no fault on [her] part … that activity creates a threat or causes harm to which the victim is in no way liable" (McMahan, 2004, 723, cited in Benbaji, 2021, 4). Finally, moral responsibility for posing an unjust threat is also sensitive to the degree to which the threatening person is evidence-relative permitted to pose it or whether she is also evidence-relative justified in posing it, a justified act being understood by McMahan as one that not only is permissible but the performance of which has a positive moral reason (McMahan, 2009, 43). Thus, an ambulance driver who drives consciously to the site of an accident to evacuate a victim but who is subject to a "freak event [that] … causes the ambulance to veer uncontrollably toward a pedestrian," is evidence-relative justified for posing such threat (McMahan, 2009, 166, cited in Benbaji, 2021, 4). This is because the ambulance driver "justifiably believes that she has a strong moral reason to do exactly what she is doing" (McMahan, 2009, 167, cited in Benbaji, 2021, 4). The conscientious ambulance

[4] McMahan develops the moral responsibility account in McMahan (2005, 394–404) as well as in several other essays. Other defenders of the moral responsibility account include Gordon-Solmon (2018), Otsuka (1994). It should be noted that McMahan's understanding of the notion of moral responsibility is different than the general concept of moral responsibility. For a discussion of this issue, see Sartorio (2021).

[5] This is a paraphrase of the criterion as introduced in McMahan (2005, 394).

driver, then, is less morally responsible for posing her threat than the conscientious car driver, for the latter supposedly does not justifiably believe that she has a moral reason to drive, given that driving a car is morally indifferent in most circumstances.

If one adopts McMahan's view, it seems straightforward that those combatants who are ordered to fight in an unjust war are morally responsible for their wrongful acts (first and foremost, killing and maiming innocent persons without justification), even if their acts presumably meet relevant in bello norms.[6] This is because the combatants both not only are *moral agents* but also can *foresee* that obeying their orders to fight might create threats of harm to innocent persons and also because, even if they *justifiably believe* that they are fighting in a just war and so are evidence-relative justified in acting as ordered (and thus are less morally responsible than if they were not so justified), obeying their orders, as a matter of fact, involves violating their opponents' rights to life and personal integrity.

Benbaji grants—though only for the sake of the argument—that these unjust combatants are morally responsible for their wrongful acts in regular authority cases.[7] But he submits that things are fundamentally different in special authority cases. If this is correct, at least some unjust combatants are not morally responsible for wrongfully killing and maiming their enemies and, as a result, are permitted to do so.

3 Regular and special authority cases

Call "Expert" a person who exercises (practical) authority over another person and "Assistant" a person over whom Expert exercises her authority.[8] Benbaji (2021, 5) argues that, in regular authority cases, the grounds of Expert's authority over Assistant lie in Expert's expertise vis-à-vis Assistant over a particular domain of action. Thus, Assistant is duty-bound to obey Expert's orders (provided that they are within Expert's area of expertise) because so acting allows Assistant to fulfil his independent duties.[9] As even experts make mistakes, however, Expert can make mistakes. In particular, Expert might issue orders obedience to which involves Assistant acting in ways that not only do not involve fulfilling his independent duties but which involve violating them.

Consider *Killing*. Suppose that Victim is drowning and that Assistant and Expert can rescue him. Suppose also that Expert cannot perform the rescue operation by herself and that Assistant can, but does not know how to do it. As Assistant is duty-bound to rescue Victim and obeying Expert's orders allows him to fulfil this duty,

[6] "Presumably" because, for a revisionist like McMahan, orders to fight in unjust wars can satisfy in bello norms only in exceptional circumstances. For a discussion, see McMahan (2009 15–32).

[7] Elsewhere, however, Benbaji and Daniel Statman explain Walzer's traditional view in terms of transferred responsibility. See Benbaji and Statman (2019, ch. 5, esp. 130–1).

[8] We follow Benbaji in using feminine pronouns when referring to Expert and using masculine ones when referring to Assistant.

[9] Here Benbaji relies on Jonathan Quong's (2011, 126–31) duty-based conception of legitimate authority, which blends John Rawls's natural duty theory with Joseph Raz's service conception.

Assistant is placed under Expert's authority in relation to how to rescue Victim. Now the moment comes at which Expert issues an order to rescue Victim. Under stress, Expert orders Assistant to hold Victim's head above the water in the wrong way. Assistant double-checks the order and suspects that Expert has made a mistake. As the order is within Expert's area of expertise, however, Assistant acts as ordered and, as a result, breaks Victim's neck, thus wrongfully killing him (Benbaji, 2021, 5–6).

Some philosophers, notably David Estlund (2007), Jonathan Parry (2017), and Massimo Renzo (2019), argue that, if certain background conditions are met and Expert makes an "honest mistake," Assistant is obligated, and therefore permitted, to act as ordered. Benbaji (2021, 6–7) replies that their arguments only show that Assistant is merely *evidence-relative* permitted to obey the order.

Special authority cases are characterized by two additional features. First, there is an explicit agreement between Expert and Assistant to a particular division of labor. Second, Superior's authority is "costly," in the sense that Assistant is incapable of determining whether Expert's orders are correct and Expert does not expect Assistant to double-check and correct her mistakes, if they occur (Benbaji, 2021, 3, 7). Thus, Assistant is duty-bound to obey Expert's orders (provided that they are covered by their agreement, which presumably are within Expert's area of expertise) both because Assistant is contractually obligated to do so and because Assistant is unable to correct Expert's mistakes, if they occur.

Consider *Special Killing*. Suppose again that Victim is drowning and that Assistant and Expert can rescue him only if they act together. Assistant and Expert can succeed in their rescue operation, however, only if they explicitly agree that Assistant will swim towards Victim and hold his head above water level and that Expert will fetch the boat that they will use to bring Victim to the shore. In addition, Assistant and Expert can rescue Victim only if Assistant obeys Expert's orders come what may—if Assistant is capable of second-guessing her orders, Expert will be unable to employ her expertise properly. So, they also agree that Assistant will take a "blurring pill," which will make Assistant unable to examine the merits of Expert's orders.[10] Now the moment comes at which Expert issues an order to rescue Victim. Under stress, Expert orders Assistant to lift Victim onto the boat in the wrong way. As Assistant is under the effects of the blurring pill and so does not realize that Expert has made a mistake, he acts as ordered and, as a result, breaks Victim's neck, thus wrongfully killing him (Benbaji, 2021, 8).

Benbaji (2021, 7–12) claims that, given that Assistant is under Expert's authority and also under the effects of the blurring pill, he (Assistant) is released from his duty to deliberate about the merits of the act that he is ordered to perform. So, Assistant permissibly lacks a necessary capacity for moral responsibility and, as a result, he is not morally responsible for wrongfully killing Victim. Thus, Assistant is not merely evidence-relative but also *fact-relative* permitted to obey Expert's order.

[10] In another version of *Special Killing*, Assistant will lose sight of the rescue mission as a whole and so he will be unable to assess Expert's orders (Benbaji, 2021, 8). Benbaji grants that this alternative version is more realistic.

4 Against "transferred responsibility"

Special Killing is highly suggestive. Depending on how one fills some missing blanks, however, it may be the case that, for reasons *other* than those put forward by Benbaji, Assistant does not wrongfully kill Victim when he breaks his neck. First, Victim may be at fault for drowning and this may imply that he has forfeited his right not to be killed by a reasonable mistake in the context of a rescue operation. Second, Victim may have forfeited this right simply by engaging in a risky activity such as swimming, even if he is not drowning for some fault of his own. If Victim has forfeited his right not to be killed by a reasonable mistake in the context of a rescue operation for either one of these two reasons, thus making himself *liable* to being so killed, then, his right not to be killed is not violated, independently of whether both Assistant and Expert are morally responsible for killing him or whether only Expert is so responsible. For the sake of the discussion, then, one must grant a key feature of *Special Killing* that Benbaji has not made crystal clear, namely that Victim has not made himself liable to be killed by a reasonable mistake in the rescue operation. We will hold this assumption in what follows.

To properly analyze *Special Killing*, it will be useful to highlight five moments in the series of events involving Assistant, Expert, and Victim. At t_1 Victim is drowning. At t_1, then, Assistant and Expert acquire the duty to rescue him, which is a duty that they can fulfill only if Assistant blindly obeys Expert's orders in relation to how to act during the rescue operation. At t_2 Assistant and Expert enter into their particular agreement, which includes the clause that Assistant will take the blurring pill. At t_3 Assistant takes the blurring pill and, at t_4, Expert orders Assistant to lift Victim onto the boat in the wrong way. At t_5, finally, Assistant obeys Expert's order and, as a consequence, breaks Victim's neck.

A natural reaction when one considers *Special Killing* is that, contrary to what Benbaji suggests, Assistant is *not* duty-bound to act as ordered. Given that obeying Expert's order involves killing a person who has not made himself liable to be killed, it is reasonable to consider that the agreement entered into between Assistant and Expert ceases to be binding at t_5 or, perhaps more plausibly, that it is just not binding at t_2. If this is correct, there is no such thing as transferred responsibility for the wrongful killing of Victim from Assistant to Expert. This is because there is no binding agreement to release Assistant from his duty to deliberate about the merits of Expert's order when receiving it.

Benbaji (2021, 9) argues that such an understanding of the case is mistaken because it assumes that Assistant's normative situation at t_2 is determined by facts that take place later on. But the future, he claims, is indeterminate, and so it cannot affect one's normative situation in the present. Needless to say, the problem of how the future affects one's normative situation in the present deserves a careful discussion of its own. We can, however, put it aside here.[11] Let us then assume that,

[11] Notice, however, that if Benbaji is right, both the conscientious car driver and the conscientious ambulance driver discussed by McMahan are probably fact-relative permitted to drive, even though they will end up killing innocent persons, unless those persons kill them first. The drivers' liability to defensive killing, then, may be affected as well.

when Assistant and Expert enter into their agreement, there is no fact of the matter as to whether Expert will order Assistant to lift Victim onto the boat in the wrong way, so that Assistant's obedience to this order will lead him to wrongfully kill Victim. This, however, does not seem sufficient to allow us to consider that Assistant's moral responsibility for the wrongful killing of Victim is transferred to Expert. This is because things are *fundamentally* alike in *Killing*. That is, when Assistant submits to Expert's authority, there is no fact of the matter as to whether Expert will order Assistant to hold Victim's head above the water in the wrong way, such that Assistant's obedience to this order will lead him to wrongfully kill Victim. Yet, Benbaji grants that, in *Killing*, Assistant is merely evidence-relative permitted to act as ordered. This verdict, then, goes against Benbaji's own analysis of *Special Killing*—it suggests that the agreement entered into between Assistant and Expert is not binding.

Nevertheless, Benbaji insists that the agreement entered into by Assistant and Expert in *Special Killing* is binding and that this accounts for his different analyses of the two cases. To make it clear that the agreement entered into between Assistant and Expert plays the momentous role of releasing him from his duty to evaluate the merits of the act that he is ordered to perform, Benbaji introduces a third case, *Semi-Special Killing*.

In *Semi-Special Killing*, Assistant can again rescue Victim from drowning. But because Assistant cannot perform the rescue operation by himself, he has to rely on Robot (a robot) as Robot is an excellent instrument for rescuing people from drowning and because Assistant has no other means for rescuing Victim. As in *Special Killing,* in *Semi-Special Killing* Assistant also needs to take a blurring pill because—the example goes—Robot cannot serve people who are able to second-guess its instructions. Finally, things also turn out bad in *Semi-Special Killing*: a freak event causes an electrical short-circuit and, as a result, Robot issues a wrong instruction. As Assistant is under the effects of the blurring pill and so does not realize that Robot has made a mistake, he acts as instructed and, as a result, breaks Victim's neck, thus wrongfully killing him (Benbaji, 2021, 9).

Benbaji (2021, 9) claims that, given that the event that will cause Robot's failure is indeterminate at the time when Assistant decides to rely on Robot to rescue Victim, he (Assistant) is under a duty to use it to perform the rescue operation. Thus, Assistant is duty-bound to rely on an instrument that, to function properly, requires that he suspends his deliberative capacities when following its instructions. Yet, Benbaji (2021, 9) also claims that, given that following Robot's mistaken instruction causes Assistant to kill Victim, who is a person who has not made himself liable to be killed, when Robot issues that instruction, he (Assistant) acquires a duty *not* to act as instructed.

This difference between *Semi-Special Killing* with *Special Killing* is, Benbaji thinks, revealing. The normative force of the agreement entered into between two moral agents is such that, in *Special Killing*, it releases Assistant from his duty to examine the merits of Expert's order. And such force is made clear when one considers the following imaginary dialogue between Assistant and Victim in which Assistant explains his behavior:

> "I was under a fact-relative duty to enter an agreement that subjects me to Expert's authority and fixes a division of labor between us. As part of this agreement, I had to agree to suspend my deliberative capacities while acting under her authority. In return, Expert released me from my duty to check the soundness of her instructions" (Benbaji, 2021, 10).

It is not completely clear, however, that Benbaji has really addressed the issue at stake, namely whether the agreement entered into by Assistant and Expert is binding. Moreover, it seems that not only his first two cases but actually all three of them are, from a normative point of view, structurally the same. In *Killing*, *Semi Special Killing*, and *Special Killing*, first, Assistant is under a duty to rescue Victim, which is an act that he cannot perform alone. And in all three cases, second, Assistant is under a duty to take reasonable means to that end, which (depending on the case) may or may not involve relying on an authority, which (also depending on the case) may or may not be grounded in an agreement that includes a particular division of epistemic labor. In *Killing*, Assistant is placed under Expert's authority because obeying Expert's orders is the only way in which Victim can be rescued; in *Semi-Special Killing*, he is placed under a duty to use Robot because things are set up in such a way that doing so is the only way in which Victim can be rescued; and, in *Special Killing*, he must enter into his particular agreement with Expert because, as the example indicates, this too is the only way in which Victim can be rescued. Now, Benbaji claims that in both *Killing* and *Semi-Special Killing* Assistant is not duty-bound, and so he is not fact-relative permitted, to obey Expert's mistaken order. As things are fundamentally the same in *Special Killing*, it seems that Benbaji is forced by his own reasoning to also admit that the fact that there is an explicit agreement between Assistant and Expert to a particular division of epistemic labor cannot play the role that he (Benbaji) believes that it does.

The reason why the agreement entered into between Assistant and Expert is not binding should be clear. This agreement allegedly binds Assistant to act in a way that is impermissible, namely it obligates Assistant to obey an order the following of which involves killing a person (Victim) who has not made himself liable to be killed by a reasonable mistake, exactly as it happens in the other two cases. That a moral agent can be obligated so to act, however, is difficult to believe. This is because, in the case at hand, being so obligated is tantamount to saying that Assistant is duty-bound to act impermissibly. But that there can be such a duty is extremely implausible.

Suppose, nevertheless, that we are wrong and that the agreement entered into between Assistant and Expert is in effect binding.[12] Even if this were the case, we will now argue, this does not entail that Assistant is not morally responsible for wrongfully killing Victim. Consider the following case. Driver* is having a drink

[12] Here is one reason why one may think that the agreement is binding. The agreement is binding because, in fact, it does not obligate Assistant to act in a way that is impermissible. This is because entering into the agreement is the *only* way in which Victim can be saved, even if he ends up being killed as a result of it. At t_2, then, Assistant engages in a risky activity that he is plausibly permitted to engage in, namely trying to save a person who otherwise will die. It is not clear, however, that Benbaji himself can avail himself of this reason and maintain his view intact.

with some friends at a bar (t_3*). Suppose that, even though Driver* is intoxicated and so should not drive his car home, he still does so and, while doing so, hits and harms Pedestrian* (t_5*). There is no doubt that Driver* is morally responsible for this wrongful act. And this is the case even though, when Driver* hits Pedestrian* with his car, Driver* is incapable of driving properly because he is intoxicated. The reason is straightforward: at t_3*, when Driver* has the drink, he has all his cognitive capacities functioning normally and so is morally responsible for his acts. At t_3*, then, not only can Driver* control his act of having the drink, but also can foresee the likely results of that act. And one such foreseeable act, precisely, is hitting a pedestrian while driving home. Driver*'s moral responsibility for wrongfully hitting and harming Pedestrian*, then, can be "traced back" to his act of having the drink.[13]

Nothing of substance changes in a case in which the driver has the drink for a good (perhaps even required) cause. Consider the following variation of the above case. Suppose that a bar customer has just had a heart attack and so needs to be taken care of quickly ($t_1^{\#}$). Suppose also that Driver$^{\#}$ is the only driver available and also that he is slightly overwhelmed by the situation and so has a drink to pull himself together ($t_3^{\#}$). Suppose finally that, while rushing to the hospital, Driver$^{\#}$ hits and harms Pedestrian$^{\#}$ ($t_5^{\#}$).[14] It is clear that Driver$^{\#}$ is also morally responsible for this wrongful act. And this is true even though, when Driver$^{\#}$ hits Pedestrian$^{\#}$, Driver$^{\#}$ is incapable of driving properly because he is intoxicated. This is because when Driver$^{\#}$ has the drink, he has all his cognitive capacities functioning normally and so is morally responsible for his acts. At $t_3^{\#}$, then, not only can Driver$^{\#}$ control his act of having the drink, but also foresee the likely results of that act, one of which is hitting a pedestrian while driving to the hospital. Driver$^{\#}$'s moral responsibility for wrongfully hitting and harming Pedestrian$^{\#}$, then, can also be traced back to his act of having the drink.

As we will now show, things are fundamentally alike if the case includes a promise. Suppose that, right after the bar customer has had the heart attack at $t_1^{\#}$, Driver$^{\#}$ quickly promises his friends that he will have a drink to pull himself together, for otherwise he would be unable to drive ($t_2^{\#}$), and then proceeds to have the drink ($t_3^{\#}$). Even if one thinks that this promise is binding, the analysis of the case is basically the same. This is because, even though Driver$^{\#}$ is supposedly duty-bound by his promise, he still has all his cognitive capacities functioning normally when having the drink and so is morally responsible for his acts. So, at $t_3^{\#}$ not only can Driver$^{\#}$ control his act of having the drink, but also can foresee the likely results of that act, one of which is hitting and harming a pedestrian while rushing to the hospital as a result of driving his car badly because of being intoxicated. Thus, Driver$^{\#}$'s moral responsibility for wrongfully hitting and harming Pedestrian$^{\#}$ can still be traced back to his act of having the drink.

[13] For an illuminating discussion of "tracing," see Fischer and Tognazzini (2009).

[14] To make the case completely analogous to *Special Killing*, Driver$^{\#}$ should harm (in fact, kill) the bar customer rather than Pedestrian$^{\#}$. As nothing of substance seems to depend on this particular issue, we prefer to make the case involving Driver$^{\#}$ as similar as possible to the case involving Driver*.

As we said above, Benbaji argues that, in *Special Killing*, Assistant is not morally responsible for wrongfully killing Victim because he (Assistant) is contractually obligated to obey Expert's mistaken order and also because he cannot detect that Expert has made a mistake because he took the blurring pill. The case involving Driver#, however, shows that, even if it is the case that Assistant is so obligated, this verdict is nevertheless incorrect. This is because, when Assistant takes the blurring pill, his cognitive capacities are functioning normally and so is morally responsible for his acts. Even though the contract *binds* Assistant, it does *not* affect his cognitive capacities, which are affected by an ulterior act, namely his act of taking the blurring pill. The act of taking the blurring pill, then, is the relevant one to evaluate Assistant's moral responsibility for wrongfully killing Victim. And, when taking the blurring pill, Assistant not only has moral agency but also can foresee that, if he proceeds to take the pill and, later on, Expert issues a deeply mistaken order, he might kill Victim if he obeys it. Thus, Assistant's moral responsibility for wrongfully killing Victim can *also* be traced back to his act of taking the blurring pill, no matter that he is contractually obligated to do it because of his agreement with Expert. So, the correct analysis of the case is that, even though Assistant is incapable of determining that Expert's order is mistaken because he is under the effects of the blurring pill when he acts as ordered, Assistant is nevertheless morally responsible for that wrongful act. If one adopts McMahan's moral responsibility account, moreover, it follows that, as a result of his moral responsibility for posing his unjust threat, Assistant has made himself liable to defensive killing by Victim (or a third party) in a variation of the case in which that is the only way in which Victim's life can be saved.

In private correspondence, Benbaji granted that Driver# is morally responsible for wrongfully harming Pedestrian# but he also said that he still believes that things are fundamentally different in *Special Killing* (under the assumption that the agreement entered into between Assistant and Expert is binding). This is because, Benbaji argued, in the driver case no one but Driver# is under a duty to deliberate on the merits of his acts while rushing to the hospital. In Benbaji's own words, "[t]he promise that the driver made [to his friends] is *not* an agreement under which the responsibility for the driving can be transferred [to them]."

It is true that in the case involving Driver#, no one but Driver# is under a duty to deliberate about how to drive while rushing to the hospital. But this is irrelevant. This is because the particular promise made by Driver# to his friends does not play a significant role when considering his moral responsibility for his act of harming Pedestrian#; the key feature, instead, is the fact that he has a drink and then drives his car. The relevant feature of *Special Killing*, then, is not that Assistant and Expert enter into an agreement that involves a particular division of epistemic labor, but the fact that Assistant takes the blurring pill that makes him incapable of examining the merits of Expert's order. This is the reason why, even if his agreement with Expert is binding, Assistant is nevertheless morally responsible for wrongfully killing Victim. Despite the fact that *Special Killing* is a case in which Assistant and Expert enter into an agreement that includes a particular division of epistemic labor, then, it does not follow that that Expert is the only one morally responsible for wrongfully killing Victim.

Moreover, even if Benbaji is right in saying that the case involving Driver$^{\#}$ and *Special Killing* are fundamentally different because the content of the agreements made between the relevant parties are different, this does not entail that, while Driver$^{\#}$ *is* morally responsible for wrongfully hitting Pedestrian$^{\#}$, Assistant is *not* morally responsible for wrongfully killing Victim. Suppose that Benbaji is right—it is relevant that the promise made by Driver$^{\#}$ to his friends is *not* an agreement under which his responsibility for his acts while intoxicated is transferred from him to them, while the content of the agreement made between Assistant and Expert makes it an agreement under which his responsibility for his acts while under the effects of the blurring pill is allegedly transferred from him to her. Suppose, moreover, that the two agreements are binding. Still, nothing of substance changes under these assumptions.

A validating condition (among others) for an agreement to be binding is that, when entering into the agreement, the parties must be in control of their acts and so that they must be morally responsible for them. So, when Driver$^{\#}$ makes his promise to his friends, he must be in control of his acts—otherwise the promise would be void—and, if Benbaji is right that the content of agreements is relevant, *that* is the relevant act to trace back his moral responsibility for wrongfully hitting Pedestrian$^{\#}$. Thus, Driver$^{\#}$'s moral responsibility for his wrongful act at $t_5{}^{\#}$ can not only be traced back to $t_3{}^{\#}$ but even to $t_2{}^{\#}$. Analogously, when Assistant enters into his particular agreement with Expert, he must be in control of his own acts—otherwise the agreement would be void—and, so, his moral responsibility for wrongfully killing Victim can then be traced back to that act. It does not follow, then, that Assistant is not morally responsible for wrongfully killing Victim. What follows, instead, is that Assistant's moral responsibility for that act should not be traced back to his act of taking the blurring pill, but rather to his act of entering into his particular agreement with Expert.

5 Final remarks

Benbaji argues that those unjust combatants who obey their orders to fight in special authority cases are not morally responsible for the wrongful acts that they commit as a result of obeying their orders to fight. We showed, first, that Benbaji's own reasoning suggests that the agreements entered into between those combatants and their superior are not binding and, second, that even if the agreements are binding, those combatants who obey their orders to fight are nevertheless morally responsible for their wrongful acts.

In Benbaji's account, if one assumes that special authority cases are sufficiently widespread (which is a claim that, as we indicated, he does not commit himself to), the traditional view that just and unjust combatants are moral equals in the sense that they are permitted to kill one another turns out to be correct. But Benbaji's account fails. Thus, there is no reason to believe that this position is correct, even if it is the case that special authority cases are widespread.

That the unjust combatants who obey their orders to fight in unjust wars are morally responsible for their wrongful acts is a verdict that supports the revisionist view that just and unjust combatants are morally unequal. This is because these

combatants are not fact-relative permitted to act as ordered. At most, they are only *evidence-relative* permitted to do so. And this is a point that the revisionist view can accommodate without much trouble.

To say that these combatants are evidence-relative permitted to obey their orders to fight is basically the same as saying that these combatants are sufficiently justified in believing that their acts are fact-relative permissible, given their circumstances, and that they are not to blame for having such those beliefs. So, the combatants act on the basis of beliefs which, although false, if true, would make their acts justified in the fact-relative sense. Yet, this does not mean that their acts are justified and so permissible; it only means that the combatants have a fully mitigating epistemic excuse and so are blameless for the wrongful acts that they commit as a result of obeying their orders.[15]

Acknowledgements We especially thank Yitzhak Benbaji for multiple comments on previous drafts, including one as referee for *Philosophical Studies*. We would also like to thank the second anonymous reviewer for helpful suggestions.

References

Benbaji, Y. (2021). Costly authority and transferred responsibility. *Philosophical Studies*. https://doi.org/10.1007/s11098-021-01615-2

Benbaji, Y., & Statman, D. (2019). *War by agreement*. Oxford University Press.

Estlund, D. (2007). On following orders in an unjust war. *Journal of Political Philosophy, 15*(2), 213–234.

Fabre, C. (2012). *Cosmopolitan war*. Oxford University Press.

Fischer, J. M., & Tognazzini, N. A. (2009). The truth about tracing. *Noûs, 43*(3), 531–556.

Frowe, H. (2014). *Defensive killing*. Oxford University Press.

Gordon-Solmon, K. (2018). What makes a person liable to defensive harm? *Philosophy and Phenomenological Research, 97*(3), 543–567.

McMahan, J. (2004). The ethics of killing in war. *Ethics, 114*(4), 693–733.

McMahan, J. (2005). The basis of moral liability to defensive killing. *Philosophical Issues, 15*, 386–405.

McMahan, J. (2009). *Killing in war*. Oxford University Press.

Otsuka, M. (1994). Killing the innocent in self-defense. *Philosophy & Public Affairs, 23*(1), 74–94.

Parry, J. (2017). Authority and harm. In D. Sobel, P. Vallentyne, & S. Wall (Eds.), *Oxford studies in political philosophy* (Vol. 3, pp. 252–277). Oxford University Press.

Quong, J. (2011). *Liberalism without perfection*. Oxford University Press.

Renzo, M. (2019). Political authority and unjust wars. *Philosophy and Phenomenological Research, 99*(2), 336–357.

Rodin, D. (2002). *War and self-defense*. Clarendon Press.

Sartorio, C. (2021). The concept of responsibility in the ethics of self-defense and war. *Philosophical Studies*. https://doi.org/10.1007/s11098-021-01614-3

Walzer, M. (2004). *Arguing about war*. Yale University Press.

[15] The discussion of the relationship between evidence-relative justification and excuse is indebted to related remarks by McMahan. See McMahan (2009, 43, 62, 144).