Ethical Theory and Moral Practice
https://doi.org/10.1007/s10677-024-10457-w

Self-Threatening Extortionists Constitute a Problem for Utilitarians, Not Contractualists

Robert Huseby[1] · **Sigurd Lindstad**[1]

Accepted: 27 June 2024

Abstract
Johann Frick has claimed that morality requires that we (in many cases) should give in to the demands of rational agents who attempt to extort us by threatening to harm themselves (self-threatening extortionists). He has further argued that since contractualism implies that there is no such moral requirement, such cases represent a problem for this brand of moral theory. In this paper, we argue that things are quite the other way around: Morality does not require that we give in to the demands of self-threatening extortionists. Such cases, therefore, represent a problem for (act) utilitarianism, rather than contractualism. Our argument appeals to a particular understanding of the idea that rational agents have a special responsibility to take care of their own interests or welfare.

Keywords Contractualism · Self-Threatening Extortionists · Utilitarianism · Rationality -responsibility

1 Introduction

Johann Frick has argued that self-threatening extortionists pose a problem for contractualism. He presents the following case:

> *Jones*. Jones, a single man with no dependents or friends, comes to your doorstep. Brandishing a pistol, he credibly threatens to kill himself, unless you pay him $20. (Frick 2016, p. 234)[1]

[1] Frick refers to the case as *The man who took himself hostage*. We just refer to it as *Jones*, for simplicity. A similar example is discussed by White (2017a, 225). White's framing of the situation as an attempt to transfer or renounce responsibility has influenced our argument in this paper.

✉ Robert Huseby
roberthu@uio.no

Sigurd Lindstad
sigurdlindstad@gmail.com

[1] Universitetet i Oslo, Institutt for statsvitenskap, Oslo, Norway

Published online: 10 July 2024

The purported problem is that contractualism allows us to stand up to such attempts at extortion, and thereby it allows easily preventable and severe harm to befall the extortionist. For Frick, this problem can only be solved by an appeal to consequentialism.

We believe that things are quite the other way around: If contractualism permits us *not* to give in to the kind of threats Frick has in mind, then so much the better for contractualism. In our view, then, self-threatening extortionists is not a problem for contractualism, but a substantive problem for utilitarianism.[2]

It is worth noting that we are concerned with showing that we are not morally obligated to give in to self-threatening extortionists (in many cases), and with showing that this is bad news for utilitarianism. We are not concerned with defending contractualism as such, apart from the indirect defense that our arguments will provide. What we claim here is presumably compatible with a range of (non-utilitarian) theories.

The paper is structured as follows: In Sect. 2, we present Frick's arguments for why we are morally required to give in to Jones' demands. In Sect. 3 we go through some preliminary points about how cases such as *Jones* should be approached when we are to use our intuitions about it as part of a more general theoretical argument. In Sect. 4 we show how attempts to explain our intuitions about *Jones* in terms of principles that focus on (a) control over the outcome and (b) the culpability of the extortionist, are unsuccessful. In Sect. 5 we dig deeper for a rationale that can back up our intuitions about Jones, and appeal to the idea that *rational agents have a primary responsibility for taking care of their own interests or welfare*. We argue that this primacy entails that when a rational agent faces small or no costs or obstacles in taking care of their own welfare and interests, responsibility for doing so remains with that agent. In Sect. 6 we show how cases of self-threatening extortionists pose a largely overlooked objection to act (and rule) utilitarianism,[3] since on that view (given the estimated consequences), we should *clearly* give in to the demands of self-threatening extortionists in a wide a range of cases. We argue that this has deeply troubling consequences, and that any moral theory which has these implications, should be rejected. This adds credibility to the rationale proposed in Sect. 5. It also turns the tables on Frick's challenge to contractualism. Section 7 concludes.

2 Frick's Argument

Frick's main claim is that cases of self-threatening extortionists show that scanlonian contractualism has trouble providing a fully plausible account of moral wrongness.[4] Further, consequentialist notions about the impersonal badness or goodness of outcomes is required to remedy this problem (Frick 2016, 260–264).

[2] In order to focus our discussion, we discuss utilitarianism in particular, rather than consequentialism in general. We do expect, however, that much of what we say about utilitarianism will have implications for many other versions of consequentialism as well.

[3] See, however, Gustafsson (2022).

[4] Without going into too much detail, this has to do with the claim that Jones cannot give a reasonable justification for why you should give him $20 to save his life, given that he himself simply can choose not to harm himself if you keep your money. According to contractualism, the question of whether Jones could give such a reasonable justification is crucial to deciding whether it is wrong to deny him the $20. For details on contractualism, see Scanlon (1998).

As noted, our interest is with the argument that, in specific circumstances, there is a moral duty to give in to the demands of self-threatening extortionists, rather than in the more general challenge to contractualism. So, what are Frick's reasons to think that we should give in to Jones' demands and pay him $20? The answer is straightforward. By paying Jones, we prevent a very bad thing from happening, at a trivial cost. Hence,

> Nonconsequentialists too, should concur that when we can avert an awful outcome at a trivial cost to ourselves or others, and there are no other ends of comparable moral importance in play, it would be morally wrong not to do so. (Frick 2016, p. 249)

This is basically what Frick has to say in favor of the claim that we should pay Jones. As such, his argument relies heavily on his moral intuitions about the Jones case.

3 The Jones case: Some Preliminaries

The *Jones* case should be approached with some care. One important premise is that Jones' mental health must be such that we can treat him as a rational and responsible agent. This can be tricky to keep in mind, as our first reaction to people who show up at our doorstep and threaten to blow their brains out lest we fork over twenty bucks, typically is that they have *some* sort of mental health issue. Frick suggests that we might imagine that Jones acts, not out of a self-centered concern for himself, but out of a concern for others, such as a church or an organization. We are not sure this helps much, as Frick's conclusion applies with equal force to purely self-interested agents. In any case, the challenge is to hold on to the idea that Jones is not mentally ill. If he is suicidal or for some other reason lacks capacity for rational judgment, it might well be true that we should pay in order to prevent him from killing himself.

We think that there is a real worry that our ability to treat Jones as a sane and rational person, and our ability to assume that the threat is credible, are mutually incompatible. It is natural to question why mentally stable people would ever decide to kill or seriously harm themselves after it becomes clear that they are not getting what they want. If the answer is unusual stubbornness, it is tempting to suggest that "unusual" needs to be replaced by "pathological". For the sake of argument, however, we will assume that the kind of cases Frick wants to assess are possible.

Further, we must assume that even though Jones' threat is *credible*, it has to be the case that if we deny him the money, he is still perfectly free to reverse his decision, and choose not to kill himself. If not, if it was *certain* that Jones would shoot himself, we might worry that he has no real choice after all, and perhaps conclude that we, for that reason, ought to give in to his demands.[5] From the extorted person's point of view, we might think of this in probabilistic terms. If we refuse to pay Jones, there is a probability X that he will shoot himself, and some probability 1 - X that he will not. Still, and crucially, whether he shoots himself or not, is ultimately up to *him*.

Another worry when assessing our intuitions about the case is that, whenever a sane person knocks on your door and threatens to kill himself if you do not give him $20, he tends to be in some sort of desperate need, meaning that the person at the door needs the

[5] This seems to be the type of case discussed by Gustafsson (2022).

$20 much more than you do. Or, at least, the person has a much stronger reasonable interest in the $20 than you have. You, on the other hand, may not be very desperate to keep your $20. In many such scenarios, it is plausible that you should give up the money.[6] These more common-sensical interpretations of the scenario may shape our intuitions about it, even if we are told that they do not apply. Thus, to evaluate the case properly, we must assume that Jones is in no relevant need, distress, bad health, or desperation, such that we ought to give him $20 for *that* reason.

We should lastly note that in this paper we only want to question the proposed *moral obligation* to give Jones $20. There are plenty of good reasons to give Jones $20 that are not what one would ordinarily think of as moral, but instead has to do with self-interest. Consider for instance all the problems (emotional and practical) associated with having someone shoot themselves on your doorstep. Another reason may simply be the uncomfortable feeling of guilt one could easily experience if Jones does indeed kill himself. Giving up $20 may be better than risking a bad feeling of guilt afterwards. This may be true even if feeling guilt about not giving Jones $20 after he shoots himself is not warranted.

4 Why Appeal to Control and Culpability cannot Explain our Intuitions about Jones

As indicated throughout, our intuition is that we are not morally required to give Jones the $20. Intuitions require explication and theoretical backing, however. An initially promising hypothesis is that there are two features of the Jones case that might help explain why we do not have a duty to give in to his threats.

1. The fact that Jones (*ex hypothesi*) fully controls whether the potential victim (himself) will be killed or not.
2. The fact that while Jones faces a very serious threat, he is also the person responsible for creating the threat, and therefore not innocent (with respect to the relevant act of extortion).

However, pointing to these factors as explanations or justifications for our intuitions, quickly runs into problems. It will be instructive to see why. Firstly, on reflection it seems that neither control over the situation, nor responsibility for creating the threat, are by themselves sufficient criteria to support our position on the question of our proposed moral duty to pay Jones. Secondly, it is unclear why these factors would be sufficient when both are present together.

To illustrate why the first criterion on its own carries little weight, consider the following case:

[6] We tend to think that this is true even if Jones does not objectively need the money, and is desperate for some subjective reason, for instance that he just desperately *wants* 20$. Such desperation may or may not be pathological, but our main concern here is to restrict our argument to cases in which Jones is, as a matter of fact, rational and sane, and in which his actions and reasoning is not distorted by desperation, need, irrationality, mental health issues or the like. We are grateful to an anonymous reviewer for pressing us to clarify this point.

> *Jones and Smith*. Jones comes to your doorstep together with Smith. Brandishing a pistol, Jones credibly threatens to kill Smith, unless you pay Jones $20.

In this case Jones surely controls whether harm will be done, but it seems that we nevertheless have a moral duty to pay him.[7] As the example is designed, the stakes are exactly the same as in *Jones*.[8] You are asked to pay $20 in order to prevent the (possible) loss of a life. And here too, it is completely up to Jones whether the harm will materialize in case you fail to give in to the extortion. Therefore, the fact that Jones controls the outcome, cannot be a sufficient condition for concluding that we do not have a moral duty to pay Jones. Moreover, by itself, the criterion seems close to irrelevant.

Much the same can be said about responsibility for creating the threat. Consider the following case:

> *Bomb belt*. Jones comes to your door and credibly claims that unless you insert a $20 bill into the complicated trigger mechanism of his straight-jacket-like bomb belt, which he has intentionally designed and strapped himself into, there is a probability that it will go off,[9] and kill Jones (but not you). The bomb belt is already activated by Jones himself, and your inserting the money is the *only* way to prevent (with certainty) Jones's death.[10]

In this case, it seems to us that we should pay. To be sure, Jones is responsible for creating the threat and therefore not relevantly innocent. There are of course plenty of reasons to criticize Jones, but we should not let him die (or allow him to be subjected to some substantial probability of dying) if we can prevent this by giving up a moderate sum of money.[11] The reason is that in this case it is not up to Jones whether he will be harmed or not if we refuse to give in. Responsibility for creating (and sustaining) the threat is thus not sufficient to conclude that we should leave Jones to his own devices. It seems, however, that responsibility is more relevant, since we would be less reluctant to pay in *Jones and Smith* than in *Bomb belt* and we would also be willing to pay more in the former than in the latter. We assume that these intuitions are not very controversial.

Our judgments about these two cases call into question the suggestion that Jones' control and responsibility in *Jones*, can support our claim that we are not morally obligated to give in to his demands. To this objection, one could respond that the two factors perhaps become morally relevant when they appear together. However, one problem with this view is exactly that each of the two factors seem relatively insignificant (though to different degrees) by themselves. This is well illustrated by *Jones and Smith*: The fact that Jones himself controls whether or not he will kill Smith has very little bearing on the question of whether we

[7] This case is structurally (if not psychologically) similar to Cohen's Kidnapper case, in which Kidnapper demands money from the parents in order to let a child go (Cohen 1991).

[8] Depending on the assumption that the probability that Jones will kill Smith in *Jones and Smith* is the same as the probability that Jones will kill himself in *Jones*.

[9] We assume again that the probability that Jones eventually dies, is the same in *Bomb Belt* as in *Jones* and *Jones and Smith*.

[10] We are grateful to Kasper Lippert-Rasmussen for suggesting the structure of this case.

[11] We do not insist that we should pay regardless of the probability. Perhaps, if the probability of the belt going off is sufficiently small, we might reasonably refuse to pay.

should spend $20 on saving Smith's life. Thus, it adds little or no weight to any claim that we should not pay. It is therefore unclear why it would add weight to such claims in the Jones case.

This means that if we want to claim that the two conditions are necessary and jointly sufficient for removing a moral duty to give into extortion, we should also explain why the presence of the two conditions *together* produce this result. However, we have no such explanation to offer. Instead, we believe that initial intuitions about the relevance of control and culpability for our judgments about *Jones*, point to a more fundamental rationale, which explains why control is important in some cases and not others. It also explains why it is natural to turn to ideas about culpability in explaining our proposed lack of duty to pay Jones. It is to this rationale we now turn.

5 We all have a Special Responsibility to Take Care of our Own Interests

Consider the fact that in *Jones*, the reasons Jones has to not kill himself in the face of adversity (the failure of others to give in to his extortion) are not simply reasons he has towards any potential victim of his actions, but also, importantly, reasons he has to take care of his *own* welfare, or interests. This is not so in *Jones and Smith*, where Smith is the only potential victim.

This observation is important because we think that all rational agents, including Jones, have a special kind of responsibility to take care of their own interests or welfare. Our suggestion is that this explains why there is no moral responsibility to give into Jones' threat.

The idea that there are good reasons to assign a special responsibility to people to take care of their own interests is a familiar one. Some of these reasons are clearly instrumental. For example, people will tend to be better at taking care of their own interests than others are, simply because they know more about those interests than others do. Moreover, they will often be best placed to take care of those interests. There are possibly also more principled reasons favoring self-responsibility. For example, under certain circumstances, there may be reasons of desert, or related reasons, to think that people should themselves bear whatever burden that results from their own choices. This may explain why we initially turned towards Jones' culpability in creating the threat as counting against a proposed duty to pay him.

The idea about self-responsibility we appeal to in this paper is not about desert, nor is it instrumental in a utility-maximizing sense. Our suggestion is that there is a general principled presumption in favor of self-responsibility that attaches to all rational agents *qua* rational agents. One might see this presumption in favor of self-responsibility as the other side of the coin of a presumption in favor of liberty.[12] Both the presumption in favor of liberty and in favor of self-responsibility may ultimately be grounded in in our capacity for rational agency and autonomy.

[12] On the presumption in favor of liberty, see Mill (1963; vol. 21: 262): "…the burden of proof is supposed to be with those who are against liberty; who contend for any restriction or prohibition… The *a priori* assumption is in favour of freedom…"; Benn (1988, 87): "…the burden of justification falls on the interferer, not on the person interfered with."; Feinberg (1987, 9): "…liberty should be the norm, coercion always needs some special justification."

Before we unpack this idea further, let us emphasize that our understanding of self-responsibility does not in any way rule out that we have demanding responsibilities towards others, to the extent that they are *in need* of our assistance. Some people are, for various reasons unable to take care of their own interests, and they may have strong and legitimate claims to help. And even if there is no lack of self-preservation yet, we have moral duties towards each other which concern (roughly) the protection and promotion of everyone's welfare and interests. However, it is our contention that in the case of (rational) self-threatening extortionists such as Jones, responsibility to protect his interests or welfare from the threat at hand, remains exclusively with the extortionist himself.

A presumption in favor of self-responsibility is supported by a particular view of the relationship between different relevant criteria for deciding when responsibility remains with the person whose welfare or interests are under threat. Naturally, we accept that the stakes matter. The extent of the potential harm,[13] and the cost of preventing it for others is, of course, relevant for deciding to what extent responsibility to take care of a person's welfare or interests is shared by other people. We submit that a further important factor are the *obstacles and costs* that are facing a person in protecting their own welfare or interests.[14]

We assume that appeals to these factors are uncontroversial. The explanation for why there is no duty to pay Jones is not found in these factors themselves, but comes down to how they relate to each other and how they are weighed against each other. Thus, the following claim is crucial to our position that there is no duty to pay Jones:

> It is only when a person faces (a sufficient level of) costs and obstacles in taking care of themselves that departures from self-responsibility can be justified.[15]

Thus, it is our contention that the following factors: (i) the costs and obstacles to a person to take care of their own interests, (ii) the extent of the harm that may befall them, and (iii) the costs to other people of helping them avoid that harm, relate to each other in a manner which means that when factor (i) is zero (or very low), the values of factor (ii) and (iii) cannot change the fact that responsibility to take care of their own welfare and interests remains with the person whose interests are under threat. A useful analogy here may be the multiplying of any number with zero. No matter how large the number is, the answer will still be zero.

Remember that in *Jones*, because Jones easily controls the outcome, *nothing* prevents him from not killing himself. On our view, this fact makes the cost of giving in to his threat or the awful nature of the consequences, irrelevant.

For the purposes of the present paper, we would like to leave open further questions about why a lack of costs and obstacles has this kind of importance for judgments about self-responsibility. However, we think that the noted analogy to the presumption in favor

[13] We intend "extent of the potential harm" here to refer to considerations about both the magnitude of the harm *and* the likelihood that it will occur.

[14] Perhaps desert, or similar ideas which attaches relevance to culpability in creating a threat, matter as well. For the purposes of this paper, we put this issue to one side, because we do not think it is necessary to appeal to it, and it would obscure our main point.

[15] Note that this formulation encompasses threats to their welfare that are not caused by the agent themselves. This seems reasonable to us. If a rational agent can easily and without cost avoid a substantial harm not caused by themselves (step aside to avoid a runaway trolley, say), the responsibility for doing so remains with the agent.

of liberty is a helpful one.[16] According to the presumption in favor of liberty, the burden of justification is always on the would-be interferer. Justifying interference by simply stating that you will interfere is pointless. Doing so constitutes no justification for interference.

Similarly, the idea that there is a presumption in favor of self-responsibility to take care of one's own welfare or interests, means that the burden of justification is always on those who want to transfer such responsibility away from themselves. *They* must point to some relevant and sufficient reason why the responsibility is not theirs. It seems to us that the fact that they refuse to take on this responsibility out of spite, say, or in order to extort other people, is simply not such a relevant or sufficient reason. This is, of course, exactly what Jones is doing. He wantonly refuses to take on the responsibility to take care of his own welfare, threatening instead to deliberately undermine it, in an attempt to transfer the responsibility onto someone else. Since he faces no obstacles or costs whatsoever in protecting his life from the gun pointing at his head, the relevant responsibility is only his.

Note that facts about costs and obstacles can change quickly as a situation develops. Consider:

> *Jones is Harmed*. Jones, a single man with no dependents or friends, comes to your doorstep. Brandishing a pistol, he credibly threatens to kill himself, unless you pay him $20. You decline and close the door. Seconds later you hear a gunshot and open the door again. Jones has shot himself but is not dead. He needs assistance to save his life.

What we have argued above, does not imply that we do not have a moral duty to help Jones at this point. Lying badly injured at your doorstep, there are now great obstacles for Jones to take care of his own interests, obstacles that were not present before he shot himself. It is therefore no longer reasonable to say that he alone has responsibility to take care of his own interests and welfare. The focus on obstacles and costs also explains why it is important to stress that our view only applies to cases where Jones is not mentally ill. Without ordinary rationality, the obstacles he faces seem to be extensive for obvious reasons.

Note also that we do not mean to imply that it would be *wrong* to give in to Jones' demands. It is not morally forbidden to take on a small cost in order to avert a great harm to Jones, even if the harm is threatened by Jones himself, and he can easily avert the harm to himself at no cost. The point is rather that it is *permissible* not to take on this cost, because the responsibility to avoid the harms, as we have argued, lies with Jones. This might be even easier to see in cases in which the harm is material.

> *House*. Jones, a single man with no dependents or friends, threatens to burn down his (valuable) house (at no risk to others), unless you pay him $20.

Here, again on the assumption that Jones is fully rational, we are not obligated to give in to the extortion. True, the cost is small, and the harm, or rather material loss, might be great.

[16] For an account of self-responsibility of the kind we are after in this paper, see White (2017b). White argues that there is a relationship between having authority over decisions that concerns one's own life and having responsibility for how well one's life goes (White 2017b, 240–242). This argument is similar to our suggestion that a presumption in favour of liberty can help illustrate the appeal of a presumption in favour of self-responsibility. See also White (2017c).

Still, it is not our job, morally speaking, to see to it that Jones' house is safe from this his own threat of destruction, when he can do so himself at no cost and with no obstacles.

One possible objection to what we have argued here is that our view of self-responsibility is too harsh on self-threatening extortionists. Suppose the stakes are even lower for us, and even higher for Jones:

> *Sharks*. Jones threatens to throw himself into shark infested waters, something which is very likely to be fatal, unless you perform a silly dance for him.[17]

Isn't it too callous to resist dancing and letting Jones take the highly dangerous jump? While we see the intuitive pull towards giving in to Jones in this case, we nevertheless think that such a criticism is mistaken. Our rationale is simple: Even if morality allows us to stand up to Jones' threat, Jones is still perfectly free not to harm or kill himself. By refusing to do the silly dance, we do nothing to increase the risk of harm to Jones. At least not in a way that is not *completely* under his own control. Assuming again that Jones is not in any way mentally ill, how can it be overly harsh on him to let whatever happens be up to him to decide? We submit that while the moral permissibility of choosing not to dance may be associated with bad consequences for Jones, this permissibility is not in fact harsh on anyone.

According to a similar objection, the reason we have given for why we have no moral duty to give Jones $20, is susceptible to the following counterexample:

> *Silent Jones*. We happen to know that there is some probability that Jones will kill himself if we do not give him $20. Importantly, however, whether or not he kills himself if we do not pay, is completely up to him, He hasn't said anything, and has not threatened us that he will kill himself, we simply happen to know it.

In *Silent Jones*, Jones seems to be solely responsible for bringing about the threat to himself, and he is free to remove that threat. What we have said about self-responsibility above should apply with equal force to the *Silent Jones* case as the original *Jones* case. Therefore, according to our view we have no moral duty to give Silent Jones $20. We could even change the cost from $20 to simply having to say a kind word to Jones.[18] According to the objection, this conclusion is highly counterintuitive, even more so than our conclusions about the original *Jones* case.

We do not think, however, that this example exposes a problem for our view. A natural thing to ask when thinking about Silent Jones, is why we would know that he is likely to kill himself if we do not give him something or say the right thing, or so on. In the example we may simply and magically know this, but in real life our suspicion about Jones would seemingly have to come from some knowledge about Jones' mental health or grasp on reality. If someone is depressed or has some mental illness that makes them suicidal or otherwise heightens the risk of suicide, then *of course* we have moral duties to try to prevent this. *Of course* any plausible moral code would demand that we say something nice to Jones if that would prevent him from committing suicide, under any normal, real-world circumstances. However, if Jones is simply sitting in silence, thinking to himself that he might commit sui-

[17] We are grateful to an anonymous reviewer for suggesting this case.

[18] We are grateful to an anonymous reviewer for suggesting this case and the objection derived from it.

cide unless this or that happens in the world, and the likelihood of him actually committing suicide is only tied to his stubbornness in going through with what he promises himself, or some such thing, and not to depression, mental illness, or irrationality, then that is something else entirely. It is not our responsibility to make his wishes come through, just because he might commit suicide, or commit other self-destructive acts, if his wishes do not come true. And this is not our responsibility, we claim, because nothing, not mental illness, not depression, not irrationality, and no external forces, prevents him from simply taking care of his own welfare, and freely choose to not act self-destructively.

6 Self-Threatening Extortionists is a Problem for Utilitarians, not Contractualists

According to Frick, self-threatening extortionists present a problem for contractualism. We disagree. If contractualism tells us that there is no moral duty to pay Jones, that counts in its *favor* as a moral theory. Moreover, in our view, cases of self-threatening extortionists pose a substantial problem for another moral theory, namely utilitarianism.[19]

Since the stakes are very high in *Jones*, and immediate intuitions about it may vary considerably, it might be helpful to consider further cases, which are structurally similar to *Jones*, but in which the stakes vary. Consider

> *Neighbor*. One morning your neighbor, Jones, commands you to mow his (large) lawn. If you refuse, he credibly threatens to smash to smithereens his own brand new, and very expensive, car.

In *Neighbor*, the stakes for Jones are much lower than in *Jones* (and most likely than in *House*). (Whether or not paying $20 involves more disutility than mowing Jones' lawn, on the other hand, depends on your preferences for money and lawn mowing, respectively). In any case, the cost of an expensive new car vastly exceeds a couple of hours of garden work in terms of utility. It seems, therefore, that according to utilitarianism, you are *morally obligated* to mow Jones' lawn. This seems slightly absurd to us. If Jones wants to destroy his car just because you refuse to do his gardening, then that is only his problem. Any moral theory which implies the opposite is on the wrong track. Consider next these three cases:

> *Harassment*. Jones, a single man with no dependents or friends, comes to your doorstep. Brandishing a pistol, he credibly threatens to kill himself, unless you perform some sexual favor for him.

> *Cutting*. Jones, a single man with no dependents or friends, comes to your doorstep. Brandishing a pistol, he credibly threatens to kill himself, unless you cut off your left little finger (painlessly, under local anesthetic).

> *Forced Adoption.* a childless couple approaches a single parent and demands to take over the single parent's baby, or else they threaten to commit suicide.

[19] See also Gustafsson (2022), who argues for a similar conclusion.

In these three cases, the stakes are clearly higher for the person being extorted than in *Neighbor*. In *Cutting*, if you do not cut off your left pinky, Jones threatens to kill himself. Unless it is very unlikely that he will follow through on his threat, the disutility of not giving in to the threat exceeds the disutility of doing as Jones says, and so utilitarianism implies that we should give in. Again, we think this is a very implausible conclusion. One suggestion may be that the idea that we are morally obligated to cut off a finger or to perform sexual favors for others is implausible no matter what the ends of such acts are. Morality, we may think, is not that demanding. However, it does not seem to be true in general that one cannot be morally obligated to give a finger to save a life. More importantly however, the fact that Jones is threatening to kill himself, rather than someone else, is clearly relevant in explaining why it is implausible to hold that we are morally obligated to give our little finger to save his life. If Jones wanted your finger in exchange for not killing Smith, rather than not killing himself, then we would more readily accept that there is a moral obligation to do. We think the same is true about the harassment case.

In *Harassment*, the specifics are left more open, which makes the case compatible with stakes of many different kinds. Surely, there are violations and abuse that are bad enough to exceed the disutility of the (potential) loss of Jones' life with a wide margin, even if the probability of him offing himself is quite high. It is equally clear, however, that there are forms of harassment and sexual extortion, that are such that they are not bad enough to outweigh some chance of a loss of a life. In our view, the specifics do not matter much here, since we do not think there is *any* duty to comply with Jones' unreasonable demands, regardless of the details, for the reasons outlined above. Utilitarians, on the other hand are clearly committed to the view that the person being extorted should give in to Jones in *Cutting*, and in many cases of *Harassment*, (that is, all those cases in which the disutility of the relevant probability of the loss of Jones' life exceeds the disutility of having to perform some sexual favor for him). This seems to us clearly to be the wrong verdict in these cases.

In *Forced Adoption* too, it seems that utilitarianism would hold that the extorted individual should give in to the extortionist demands. Admittedly, the consequences here are not obvious, but supposing the child (somewhat improbably) can get as good a life with the extortionists, and supposing the loss of the baby (whose *life* is not lost) to the single parent is not as bad as the loss of *two* lives, the extortionists should get the baby. There are many further possible, and even more horrendous illustrations, but these suggestions are sufficient for us to conclude that any theory which implies that there is a moral obligation to give into the self-threatening extortionists demand in cases such as these three, should be rejected.

One might challenge this conclusion by pointing out that act utilitarians should not look myopically at only the most immediate consequences of their actions. In extortion cases, arguably, the right response is to not give in to the demands, precisely because giving in might incentivize further extortion in the future. Thus, act utilitarianism will in fact require that we should not pay Jones.

In response, we submit first that the theoretical challenge to utilitarianism remains, or so we claim, *even if* it is the case that not paying Jones would result in the best consequences, over time. The reason is that regardless of what the consequences are likely to be in our actual world, it is easy to construct Jones-like cases that are one-off, or very rare events. If we assume that this is known, utilitarianism straightforwardly implies that we should give in to Jones. To us, as indicated, this is sufficient to cast the theory into serious doubts.

Secondly, it is far from obvious that utilitarianism requires us to not give in to Jones' demands even if we take the wider consequences into account. Suppose Jones shows up at

your doorstep before you have heard of any such incidents from others. And suppose you consider the possibility that giving in will provide him with incentives to continue his extortion, and perhaps even provide others with incentives to join the action. However, it seems that, given what is at stake, and given that you do not know how many Joneses there will be, or how others will act, it is very risky, in terms of utility, to take the chance of denying Jones the 20$.[20] Further, it is not obvious that denying Jones the money *reduces* the chances that he will move on to the next house (if he does not kill himself upon his first rejection). That depends on how much he wants the money, and how much money he wants.

Third, it is worth noting that for every case of extortion, it is not clear that the mere exchange of money provides much net disutility. Sure, if you give in, you lose 20$, but Jones *gains* the same amount. In the long run, if the number of Joneses skyrocketed, and each Jones became rich, and richer than their victims, a net disutility would probably result from most transactions, due to the diminishing marginal utility of money (again, not counting the utility gain of keeping Jones among the living). But it is very hard to assess the likelihood of this extreme case.

There is also arguably an additional disutility connected with the unpleasantness and inconvenience of being extorted. On the other hand, there might be disutility connected with Jones' potential disappointment from being denied the money too. In short, since the individual cases of extortion do not alter the total amount of utility to any significant extent when we exclude the potential loss of a life, the possibility of Jones dying will carry a lot of weight for utilitarians.

Suppose that in the interest of long-term incentives you refuse to pay Jones. Suppose further that your action inspires others to do the same. Thirty people deny Jones money, and then he shoots himself. Thirty people saving 20$ is not enough to weigh up the loss of a life. Of course, the numbers and probabilities are highly unclear here, and this is but one (random) example, but the general point is that by not paying, you risk that Jones kills himself. By paying, you make sure that he doesn't. It seems that given these considerations, not paying in order to maximize utility in the long run, is a highly uncertain course of action if utility-maximization is your primary concern.

In addition, it does not seem likely that there will be many Joneses. The reason is that, since the threat of self-harm is, by hypothesis, credible, the extortion is *risky* for Jones. Not everyone is a utilitarian, and most likely everyone won't pay (even if we are right that utilitarianism requires paying). Thus, it is not clear how many instances of extortion will be avoided by choosing not to pay in order to avoid incentivizing further extortion.

We do not claim that we have now shown that utilitarianism implies that we should pay Jones even if we take long term consequences into account. Doing so is close to impossible, because long term consequences are notoriously hard to assess. What we have tried to do, is to point to a number of considerations that we think makes it likely that utilitarianism implies that we should pay Jones even when we (to the best of our ability) take such long term consequences into account.

We have so far discussed act utilitarianism, but what about rule utilitarianism? According to rule utilitarianism, we should establish and follow those societal rules that in the long run yield the best consequences.[21] It is arguably quite tricky to assess what effects different rules concerning self-threatening extortionists will have in the long run. Let us start by considering a simple case in which we knew for certain that there is only one Jones-type extortionist,

[20] The same is true, we think, of the other cases presented above.

[21] Se for instance Hooker (2002). Note that designing the rules that will yield the best consequences in the long run is different from taking into account the long term consequences of one's own individual act.

and that he will attempt extortion only once. Again, the probability that he will kill himself if he does not receive the money, is high enough to ensure that the expected consequences of refusing to pay are worse than the expected consequences of paying.

If so, it seems that rule utilitarianism would imply a rule that said that we should pay self-threatening extortionists in these types of cases, since such a rule, the one time it is applied, clearly will produce better consequences than a rule that says that we should not pay. Moreover, it seems that even if there are many Joneses who will each attempt extortion many times, this should not alter the rule, so long as the utility calculus remains the same (or relevantly similar) in all cases. Of course, if some people are extorted many times, the cost of paying Jones might become relatively higher, and at some point, become so burdensome as to outweigh some probability of Jones killing himself. If it is likely that there will be sufficiently many such cases, the rule must perhaps be that we should not pay. (Imagine a world in which a class of Joneses extort all others to the brink of starvation). Here too, it is hard to assess the likelihood of the different scenarios.

Assume first that we have reason to believe that a rule demanding us to pay will produce the most utility in the long run. This might have some counterintuitive implications. Above we indicated that the incentives provided by single acts might be difficult to gauge in the context of act utilitarianism. In the context of rule utilitarianism things are probably different. If the rule in question is public, and people in general abide by it, it is quite likely that this will incentivize more Jones-like types to go into the self-threatening extortion business. After all, if we can expect that everyone pays, the economic upside is quite substantial. This will, by itself, however, increase the likelihood that such cases will multiply, and by extension also increase the likelihood that we end up in a situation in which a rule saying that we should pay will not, after all, realize the most utility, because so many people are extorted so many times that the overall utility calculus changes.

Perhaps the reverse can be said about a rule saying that we should not pay. Surely, as noted, such a rule will increase the chances that Jones will kill himself, and that many will do so, if there are many Joneses. On the other hand, if this rule is public and people generally abide by it, Jones has reason to expect that people will not give in, which makes his extortion scheme not only risky, but downright hazardous. This, in turn, might give him incentives to take up a (different) hobby instead. All in all, it seems to us that whether or not rule utilitarianism implies that we should have rules telling us to pay self-threatening extortionists or not, is very hard to assess.

7 Conclusion

In this paper we have argued that that we should not give in to self-threatening extortionists, if they are not under pressure with regards to their rationality, if there is a probability that they will not follow through on their threat, and if it is up to them whether they do follow through or not. The rationale behind this argument is that all rational agents have a special responsibility to take care of their own interests and welfare, and that, when they can do so with no or very low costs or obstacles, they simply have no grounds on which to transfer this responsibility on to others. This conclusion, if correct, poses a serious objection to utilitarianism, because utilitarianism seems to imply that we should give in to self-threatening extortionists, at least in wide range of cases. Whether or not this objection extends also to rule utilitarianism, is not entirely clear.

Acknowledgements The authors would like to thank Didde Boisen Andersen, Jakob Elster, Sebastian J. Conte, Göran Duus-Otterström, Silje Aambø Langvatn, Fredrik D. Hjorthen, Kasper Lippert-Rasmussen, Alejandra Mancilla, Jørgen Pedersen, Attila Tanyi and two anonymous reviewers for very helpful comments on earlier drafts.

Author Contribution The authors contributed equally to this article.

Funding The authors did not receive funding for this article.
Open access funding provided by University of Oslo (incl Oslo University Hospital)

Data Availability Not applicable.

Declarations

Ethical Approval Not applicable.

Informed Consent Not applicable.

Human Participants and/or Animals Not applicable.

Competing Interests The authors have no competing interests with regards to this article.

References

Benn SI (1988) A theory of Freedom. Cambridge University Press, Cambridge
Cohen GA (1991) Incentives, Inequality and Community. In: Petersen GB (ed) The Tanner lectures on human values. Vol. XIII, University of Utah, Salt Lake City, pp 261–329
Feinberg J (1987) Harm to others. Clarendon, Oxford
Frick J (2016) What we owe to hypocrites: Contractualism and the Speaker-Relativity of Justification. Philos Public Affairs 44(4):223–265. https://doi.org/10.1111/papa.12076
Gustafsson JE (2022) Bentham's mugging. Utilitas 34(4):386–391. https://doi.org/10.1017/S0953820822000218
Hooker B (2002) Ideal code, real world: a rule-consequentialist theory of morality. Oxford University Press, Oxford
Mill JS (1963) In: Robson JM (ed) Collected Works of John Stuart Mill. University of Toronto, Toronto
Scanlon TM (1998) What we owe to each other. Harvard University Press, Cambridge MA
White SJ (2017a) On the moral objection to Coercion. Philos Public Affairs 45(3):199–231. https://doi.org/10.1111/papa.12098
White SJ (2017b) The centrality of one's own life, 7 edn. Oxford studies in normative ethics
White SJ (2017c) Responsibility and the demands of morality. J Moral Philos 14(3):315–338. https://doi.org/10.1163/17455243-46810062