

**This is a pre-publication draft. Please cite the final version.**

## **Abstract**

Many people feel that respecting a person's autonomy is not sufficiently important to obligate us to stay out of their affairs in all cases; but the ground for interference may often turn out to be a hunch that the agent cannot *really* be competent, or cannot *really* know what her decision implies; for if she were both of these things, surely she would not make such a foolish decision. This paper suggests a justification of paternalism that does not rely on such appeals. I argue that in cases where an agent will undergo a significant alteration in their evaluative outlook – 'evaluative shift' – three central, persuasive objections to paternalism lose their force, and offer a *prima facie* case for paternalism in some of these cases. I then suggest that we can extend this argument to some cases where evaluative alteration is not predictable, but where the risk and harm are both significant. In such cases, paternalism may be justified.

## **1. Introduction**

Sometimes people make choices that seem to the rest of us bizarre, dangerous, or simply irrational, yet we nonetheless feel uncomfortable interfering in their decisions. Many of us are, in other words, instinctively averse to paternalism. But when faced with extreme cases, an absolute opposition to paternalism can also seem unattractive; when a person, no matter how competently, chooses to put themselves at risk of serious harm, many people feel that they do have both a permission and an obligation to intervene.

This suggests a deep tension at the heart of our attitudes to paternalistic intervention. Respecting a person's autonomy can feel as though it is not sufficiently important to obligate us to stay out of their affairs in all cases. At the same time, it feels as though it is of the highest importance, the aspect of a person that we must respect above all others. Indeed, we may find ourselves justifying intervention not because autonomy's value is outweighed, but on the grounds that the agent in

question cannot *really* be competent, or cannot *really* be in full cognisance of the what her decision implies; for if she were both of these things, surely she would not make such a bizarre decision. Alternatively, we might point to the wealth of evidence from empirical psychology that suggests that when it comes to choosing both means and ends, we are “intractably irrational, and...this can't be rectified by simply care and introspection” (Conly 2013, p.7).

This paper outlines a justification of paternalistic action that attempts to avoid such appeals, i.e., a justification that does without the suggestion that the target of paternalistic interference has made some mistake about what is in her interests; this view thus aims to be entirely consistent with subjectivism about well-being. I suggest first that in cases where we can reliably predict that the agent will undergo a significant alteration in their evaluative outlook – an ‘evaluative shift’ – several otherwise persuasive objections to paternalism lose their force, and may provide some support for paternalism. I then suggest that we can extend this argument to some cases where evaluative shift is not predictable, but where the risk and harm *of such a shift* are both significant. This extension is vital, since cases where we can predict with certainty that someone’s evaluative outlook will change in the way my argument requires are rare.

The paternalism I am defending is, in the terms set out by Dworkin (2014), a form of hard (as opposed to soft) determinism, because it sanctions interventions<sup>1</sup> even when we know that the agent in question would rather not be interfered with, and even if she is fully aware of the outcome that we think speaks against her action. My view is that it is also a form of strong (as opposed to weak) paternalism. The paternalism I defend does not only justify interventions on the basis of a judgement that the agent has made a mistake about the *best means* to her ends. In general, we can contrast this with a kind of paternalism that judges some of an agent’s ends to be incorrect. While my argument will apply to some such cases, it also covers cases that do not fall neatly into either of these camps. This is because the distinction between means and ends-based paternalism

---

<sup>1</sup> Although paternalism may also apply to omissions, I will for simplicity discuss only interventions.

becomes somewhat hazy – as do many apparently clear distinctions in moral philosophy – when we introduce the idea of predictive uncertainty.

In particular, we might agree with an agent both that her end is desirable, and that her intended means are the best means to that end, but nonetheless wish to intervene paternalistically because we disagree with her about whether the respective *likelihoods* of success and failure justify her action. In such cases, we do not judge the agent's end to be undesirable, but we do think that that it is not sufficiently desirable given the risk of failure, and the harms that failure will entail.

One might class this as a form of weak paternalism because it is a disagreement about means. But if we define a weak paternalist as Dworkin does, as someone who believes it legitimate to interfere with some means to an end when 'those means are likely to defeat those ends', then we will face a question of exactly how high the likelihood must be of a risk paying off before we grant that the means, while not certain, are not 'likely to defeat' a particular end. And whether this judgement is itself a judgement about means, ends or something else is unclear. Many ordinary instances of paternalistic policy cover such cases. The motorcyclist who declines to wear a helmet need not think that a head injury is an attractive prospect, and may see that wearing a helmet is the best way to avoid such injuries; she might even agree on how likely she is to incur a head injury. But she may nonetheless judge it to be *sufficiently* unlikely – given the benefits of not wearing a helmet – that it is worth the risk.

The other complication to this distinction is that, as I outlined above, my argument rests on cases where someone's ends change, or may change, quite considerably. In such cases, where we protect a person's future interest by overriding their current ends, it is not so clear that we can talk in any simple way of judging the current ends to be wrong; if the fact that motivates our intervention is the alteration of a person's ends, we might still think that if they *continued* to hold those ends throughout their life, we would not be justified in intervening. Nonetheless, I claim that our intervention would be strongly paternalistic because it acts against the individual's current ends.

The paper proceeds as follows. Section 2 outlines three objections to paternalism: that it embodies a kind of epistemic arrogance; that it is contrary to the subject's well-being; and that it is disrespectful. Section 3 then considers a familiar defence of paternalism from Richard Arneson, and suggests that while it is certainly a plausible response in some ways, it does not fully account for compelling considerations on the other side of the case.

Section 4 then explains the idea of evaluative shift, and details how cases of evaluative shift seem to undermine the force of each of the objections considered in Section 2. To be clear, my claim at this point will not be that I have offered a full argument in favour of paternalism. There are other concerns about paternalistic influence – such as the level of state power required for certain kinds of paternalism – that may limit the scope of my argument. Nonetheless, engagement with the three objections that I consider is a significant result, since they comprise three of the most significant direct objections to paternalism. Undermining these objections, I will argue, provides us with a significant *prima facie* justification for paternalistic intervention, which may be overridden if other persuasive arguments against paternalism apply.

It is worth considering in a little more detail how the undermining (if I am successful) of three objections constitutes even a *prima facie* case for paternalism. After all, why not say that even where these arguments are defeated, our default position should be to avoid paternalism? Section 4 also notes that several of the considerations that I consider as objections may actually lend some force *in favour* of paternalistic intervention in cases of evaluative shift. I also argue that there is a general presumption in favour of helping to relieve significant suffering or harm when the subject of that harm is unable to do so themselves, and that this applies in cases of evaluative shift. As such, it seems warranted in such cases to say that paternalism provides the case that must be defeated. A final justifying consideration is that where we face a choice between paternalism and rescuing people after they have suffered a particular harm following evaluative shift, we are permitted to choose the option that is less costly to ourselves.

Section 5 assesses the scope of my argument. Cases where we can be certain that someone will undergo evaluative shift are few and far between. So it may seem as though, however successful my argument is in theory, it will have little pragmatic import. I suggest that we can extend the argument to more realistic cases because even the *risk* of evaluative shift may justify an intervention if the result will be sufficiently bad. Although I do not offer exact parameters for this extension, I end by suggesting further issues that need considering.

## **2. Three objections to paternalism**

I will now briefly outline three common and, in my view, broadly persuasive objections to paternalism. I will not go into a detailed defence of their persuasiveness, for my claim is only that insofar as they *are* persuasive, they in fact offer support to the kind of paternalism that I outline. As I have said, there are other potential objections to paternalism, and these may set limits on the scope of paternalistic intervention even if my argument is successful.

### **2.1 Epistemic arrogance**

The first objection is that paternalism involves a kind of *epistemic arrogance* (e.g. Tännsjö 1999, p16). Although most people are ignorant about many things that are relevant to our decisions, it is in principle usually possible for those with the relevant empirical knowledge to communicate this information to us. On the other hand, people are experts about a particular, central kind of information, namely that regarding our 'own feelings and circumstances' (Mill 2003/1859, p.148). As such, we are under a defeasible epistemic obligation to assume that people know their own preferences and values. So long as a person has had other relevant information successfully communicated to them, or has decided not to seek additional information, then their estimation of whether a particular course of action is the right one should be assumed correct, absent robust indications of serious epistemic failure. To insist otherwise is to ignore our own epistemic weakness in an arrogant way. This may not ground an objection to weak paternalism, because there

may be cases where an agent simply refuses to accept reasonable advice about what is likely to occur from a given course of action. But it does seem a powerful response to strong paternalism; if the agent knows what is in store, it seems unjustified for us to insist that it will be bad for her when she insists that, in light of her particular preferences and values, it will be good.

## **2.2 Well-being**

The second objection to paternalism is that being able to freely choose among an array of options is itself good for people, even if it leads to otherwise sub-optimal outcomes. It may be that a lack of interference is instrumentally good for people because that lack of interference itself makes them happy, and interference makes them unhappy; or it may be because autonomous choice is partially constitutive of a good life. According to the instrumental view, a general policy of paternalism is less likely to promote people's well-being than a general policy of respect for autonomy (although this leaves open the possibility that a more limited policy of paternalism with regard to specific kinds of decisions is justified). According to the constitutive view, a policy of paternalism robs people of a particularly important and central form of well-being or flourishing, which is the ability to live life according to one's own values, and to exercise control and authorship over one's life. As Mill (2003/1859, p.141) puts it, this latter view supposes that a person's 'own mode of laying out his existence is best, not because it is best in itself but because it is his own mode'.

## **2.3 Respect**

The final objection that I will consider is that paternalism expresses a lack of respect for the subject of the intervention. Competent agents should have ultimate control over their own minds and bodies, up to the point where they harm others. Paternalistic intervention fails to respect a person's evaluative judgements about her own life, and her moral status as an individual with the right to decide for herself how her life goes. For instance, Shiffrin (2000, p.213) contrasts paternalistic

intervention with persuasion, suggesting that while the latter ‘provides reasons to [an agent]...appealing to [him] to change his mind and exercise his agency in another way’, paternalism ‘manifests an attitude of disrespect towards highly salient qualities of the autonomous agent’, such as ‘the capacity of the agent to judge [or] the capacity of the agent to act’. Similarly, Waldron (2014) asks, ‘What becomes of the self-respect we invest in our own willed actions, flawed and misguided though they often are, when...our choices are manipulated to promote what someone else sees (perhaps rightly) as our best interest?’.<sup>2</sup> Paternalism is thus deemed to be both distressing and demeaning to the agent who is judged to be worthy of less than full respect, and intrinsically wrong because it fails to afford the proper respect to a competent agent in the governance of her own affairs.

### **3. A familiar defence of (some) paternalism**

This section briefly outlines a familiar defence of paternalism offered by Arneson (2005), since my own argument shares some features with this view. Arneson considers the absolutist anti-paternalism outlined by Feinberg (1984), which bases opposition to paternalism on a strong respect for personal sovereignty over individual decisions. Arneson’s response centres on a hypothetical case in which a strong respect for personal sovereignty has apparently implausible implications for how we should behave. The case concerns a teenager who over-reacts to a moderate setback, no longer sees any worth in life, and decides to kill himself. This young man’s evaluative outlook at the time suggests that his best option is suicide. The rest of us, presumably, disagree. If we are to oppose paternalism absolutely, and respect the teen’s personal sovereignty without limit, we should let him commit suicide; to prevent him from doing so is to fail to respect his sovereignty over his own decisions, body, and life. But, says Arneson, it is clear that we ought not to let him kill himself.

---

<sup>2</sup> Waldron’s discussion explicitly targets the idea of ‘libertarian paternalism’, outlined in Sunstein and Thaler (2008). *Nudge: Improving Decisions About Health, Wealth, and Happiness*, (Yale University Press, 2008). But it also seems to apply to the more traditional paternalism I discuss here.

So there are at least some limits to personal sovereignty, and some permissible cases of hard paternalism.

I agree with Arneson's diagnosis of what we ought to do: we ought to act paternalistically in this case by preventing the teen from suicide. But Arneson's explanation of his case falls quite explicitly, and without embarrassment, into the category outlined in Section 1 of insisting that the young man must have made a mistake in his evaluative outlook. He sets up the case (2005, p.263) as one where 'the good of the individual that is at stake can be enormous and the degree to which paternalistic interference would frustrate the agent's interest in self-determination can be very slight'. Essentially, then, the explanation is that the liberal right to autonomous decision-making is simply outweighed by the enormous value at stake: a young man's life. Even though the individual in question does not value his life – that, after all, is the very reason he has decided to kill himself – the rest of us know that it is better for him to live. Our paternalism is justified by the fact that the teenager has made an obvious mistake.<sup>3</sup>

This is, then, to ignore the original intuition at the heart of the anti-paternalist sentiment: that if a person is competent, then he or she has approached the same situation as we do, and made a different call. Arneson's insistence that the teenager's life is just obviously more important than respecting his current wishes is an insistence that we ought simply to ignore some central evaluative judgements. In this case, it is backed up by the implicit assumption that the teenager must simply not be competent with respect to this decision; he is making the obviously wrong choice. I do not wish here to argue directly against that view; maybe it is just objectively true that such evaluative judgements as the teenager makes are incorrect. What I am interested in here is whether the appeal that Arneson makes to a simple balance of good is the only option, or whether there may be

---

<sup>3</sup> One might justify intervention in a different way, by appealing to the *impersonal* value of the teenager's life. Bou-Habib (2011) offers a view quite like this, though he bases his assessment on the impersonal value of autonomy, not life *per se*. It is certainly true that this would no longer necessitate the judgement that the teenager was making an evaluative mistake; but, as Bou-Habib suggests, it is also not clear that this would any more be a case of paternalism, since the individual's good is no longer our motivation in intervening.

something more fundamental at stake in cases where we may permissibly override such apparently competent preferences.

#### **4. Evaluative shift**

One reason that Arneson's assessment seems persuasive is that that the teenager is only temporarily suicidal. There is something *unstable* about his evaluative stance, considered against a background of otherwise non-suicidal behaviour, which makes us doubt whether the choice of suicide is 'really' what he wants. If this is the justification underlying Arneson's reasoning, then it suggests some limits on the extension of the case for paternalism. We might consider a more stable desire to end one's life, backed up by relevant information that suggests that one's situation is not going to improve, e.g., a relevant psychiatric diagnosis of untreatable depression. This is at least reported to have been the judgement in a recent case in Belgium, where doctors granted a medically depressed woman the right to end her own life on the grounds that her mental health was unlikely to improve (Buchanan, 2015).

However, such an evaluation leaves it unclear what exactly is to count as a 'stable' preference. If stability is defined in purely temporal terms, such as 'lasting for at least amount of time  $T$ ', this leaves it open that we might have a case where we can predict a similarly radical change in evaluative stance as we might predict in Arneson's teen (i.e., we assume that he will at some point *not* want to commit suicide, and regard his previous desire as mistaken), but which is held for long enough to count as stable.

Moreover, there is an alternative explanation of the force of Arneson's original case: when we assume that things will improve for the young man, we judge that there will come a point when he will undergo a dramatic 'evaluative shift' from his current desire to end his life. And while such shifts may come as part of an unstable, fluctuating set of attitudes (as with Arneson's teenager), they may also occur in a more stable, gradual way. Consider the case, outlined by Bou-Habib

(2011), of a person with a strongly youth-oriented view of the good. While young, this person prefers to maximise gains in their youth, even at great cost to their old age. They do not strictly desire to have a meagre old age, but they approve trade-offs that most of us would think unreasonable because they involve only minor gains at one point in life for much greater losses elsewhere. If we compare this agent's evaluative outlook in her youth and in old age, we may find a quite dramatic difference; but that difference may have emerged through gradual, slow change over the course of a lifetime.

My view is that we are justified in preventing this agent from making decisions in her youth that will radically undermine her ability to live in her old age, just as we are justified in preventing the teenager from suicide. My explanation is that what unites these cases is that there is a reasonable prospect of a significant evaluative shift. To the extent that we judge the legal reaction in the Belgian woman's case to be a reasonable one, it seems plausible that this is because we agree with the judgement that she cannot expect things to improve.

It is obviously possible to insist, just as Arneson does about his teenager, that the youth-oriented agent simply makes an evaluative mistake. One might claim that it is irrational to be anything other than temporally neutral with respect to one's own good; preferring benefits simply because they occur in one's youth is thus not a rational stance. I think this oversimplifies things considerably. But even if this response is right about the rationality of the agent's behaviour, it undermines only the first argument against paternalism, offered in Section 2.1, which is the epistemic arrogance objection. If we *know* that it is objectively bad for an agent to behave in a certain way, then it is not epistemically arrogant to forcibly prevent that behaviour. But paternalism might still be ruled out on the grounds of (at least) the other two objections: the value to the agent of letting her live as she wishes, and the disrespect evidenced by paternalistic intervention, even if it is epistemically justified. And unlike Arneson's teenager, the youth-oriented individual may have a fairly stable preference for acquiring benefits through her youth, even if it ultimately changes in old age. Her

preference is not mere whimsy, or a flash in the pan over-reaction, but a genuine, deep-seated assessment of the best life.

I will now expand my own argument in favour of some degree of paternalism in the case of a person with a youth-oriented view of the good. This argument does not rely on the claim that the person's preference for her youth is irrational. Rather, it relies on the possibility that the agent will undergo a radical evaluative shift with regard to her view of the good. Such an argument might strictly justify *either* restrictions on behaviour before the agent acts *or* rescue when the agent is suffering the consequences.

Consider the position of our agent when she reaches old age. Because she had a deeply youth-oriented view of the good, she has saved nothing. She has also failed to look after her health with regard to issues that did not affect her welfare when she was young. As far as her current attitude goes, she may be in one of two states. Perhaps, although she suffers in her age, she judges the sacrifice to have been worthwhile because she still holds a youth-oriented view of the good. But it is also possible that she now no longer identifies with her past choices, because she has undergone a significant evaluative shift. If she were she given the chance to live her life again, she would not make the same choices, because they seem to her to be evaluatively mistaken.

I have said that our agent 'no longer identifies' with her previous evaluative stance. This is not merely a case of regret. I might regret a decision I have made, and yet still identify with the values that prompted that decision. For instance, if I take a risk that on balance seems to me worthwhile, I might regret my action if – against the odds – the risk fails to pay off. And yet if I reflect on the decision, I might still be inclined to say that it was the right one, given the information I had. I need not regard the evaluative outlook that prompted my decision as mistaken, and I might well take the same risk in equivalent circumstances. The agent in my case, on the other hand, now regards not just her actions but her evaluative stance as mistaken.

Evaluative shift is not merely a case of incomplete knowledge, where the agent in her youth failed to anticipate how bad things would get. We can even imagine that in her youth she anticipates perfectly well that when she reaches old age she will regret, and fail to identify with, the trade-offs she is currently making; but that regret is just one more cost that she is prepared to trade against her youthful happiness. An important element of evaluative shift, then, is the implication that regret and failure to identify need not imply that the original decision was a mistake. It is entirely consistent with this idea to say that the youth-oriented individual is acting rationally according to her present values, and that she will rationally regret what she has done when she is older (see Wallace, 2013 for related ideas).

Finally, it is worth noting that an agent who undergoes evaluative shift need not regard herself as a ‘different person’ in any deep sense. She might, for instance, still identify with a number of other evaluative stances that she held; and even if she does not, she may still regard herself as the same person. Although I will for the sake of ease sometimes talk as though the agent in the present and in the future were two different persons – such as speaking of the ‘present agent’ and ‘future agent’ – I do not intend for this to have any metaphysical implications.

I will now argue that none of the three objections to paternalism outlined in Section 2 give us conclusive reasons not to act paternalistically. We are justified in insisting to some extent that the youth-oriented agent saves and plans for her old age, even when she could use those resources to improve her youthful existence; and we may be justified in incentivising her in potentially manipulative or coercive ways, e.g., by taking some of her current earnings and refusing to release them until she is old.

#### **4.1 Epistemic arrogance**

Consider first the objection that paternalism involves a kind of epistemic arrogance. That objection rests on an asymmetry between the kind of information that an agent might lack (e.g. external

empirical information) and the kind that the rest of us seem to have reason to accept her word on (her own assessments of her preferences and values). This asymmetry is absent in a case of significant evaluative shift because an agent's current evaluative stance is unreliable evidence of their future evaluative stance, particularly when the future we are considering is relatively distant. The agent might insist when she is young that she will not regret her choices when she grows old. But this is itself a claim to external empirical knowledge, rather than to internal knowledge. The introspective advantage the agent has to her own current values – which grounds the epistemic arrogance objection – does not extend to her own distant future values. She can certainly make semi-educated guesses about what she will care about in the future; but in this sense she is not obviously better off than others with regard to her own future evaluative stances. Indeed, in some cases an agent may be in a worse epistemic position: for others might be able to observe, from a more detached perspective, that youth-oriented evaluative stances are often abandoned as people grow older, or that the agent is exhibiting various behavioural signs that are associated with some level of (potentially long-term) evaluative instability, and infer that this is also likely in the case of our agent.

In cases where an agent is in a significantly worse epistemic position with regard to whether she will undergo evaluative shift, it is epistemically arrogant of *her* to insist that she knows, better than anyone, what will be good for her. If we have good reason to suspect that she will undergo evaluative shift – say, because most people who have been in her position do change their views as they get older – then we are, unlike in the case where we claim to know her *current* values better than she does, in a better epistemic position than she is.

Of course, there will be many cases where neither we nor the agent can predict with any certainty whether she will undergo evaluative shift. In such cases, I accept that the epistemic objection may provide neither support nor opposition to paternalistic intervention. Nonetheless, there may be

some grounds even in cases of epistemic parity – on the basis of other considerations that I outline below – to justify some intervention. I discuss this issue in more detail in Section V.

On the other hand, the agent might accurately predict that she will undergo a significant evaluative shift, but insist that this shift will be a mistake on her part, and hence that her predictable future values give her no reasons now. Is it not epistemically arrogant of us to insist on respecting these future values despite her current distaste for them? I think not. For the original reasons behind the accusation of epistemic arrogance now seem to extend to the agent's own assessment of her future values: she declares that despite the apparent sincerity of her future rejection of the youth-oriented view, her evaluative shift will constitute a mistake on her part. But it is not clear why this is any less epistemically arrogant than our paternalistic insistence that, despite the sincerity of her *current* evaluative stance, she is simply mistaken about what is really good for her.

To be clear, my argument is that the considerations in favour of regarding the future agent as the better judge of future interests than the present agent are *the same considerations* as those in favour of regarding the present agent to be the better judge of her own *present* interests than we are. Just as we lack first-personal insight into an agent's current evaluative stance, so do all of us lack first-personal insight into our future evaluative stances. The asymmetry between each of us and all other people regarding our present values is matched by the asymmetry between my current and future 'selves' regarding my future values. So it stands that if the epistemic arrogance argument is a successful complaint against ordinary cases of paternalism, then this is a successful response.

## **4.2 Well-being**

The second objection I outlined is that we should oppose paternalism because living according to one's own view of the good is either instrumentally or intrinsically good for a person. But it is questionable whether respecting our agent's wishes to sacrifice her old age for her youth really can be construed in any simple way as respecting her ability to live according to her own view of the

good. After all, the agent that we are considering has two very different views of the good life at different times. If we follow her youthful view of the good, and allow that to dictate the course of her entire life, it seems clear that we are *failing* to allow any latitude for her view of the good when she is older. A person whose youthful choices now leave her in a position that, according to her own view of what is best, is very badly off may rightly complain that we *did nothing* to try to prevent this occurrence when it was in our power at least to limit its effect.

This has two implications. One aspect of the well-being concern was simply that people do not like being meddled with. The facts of evaluative shift do not change this; a person who undergoes paternalistic intervention may still resent it at the time it happens. My suggestion is simply that, if evaluative shift does occur, this person may later be *glad* that we intervened. As such, there is a balance to be struck between present and future costs, and hence no straightforward argument against paternalism. So I do not argue that all cases of evaluative shift justify paternalism; rather, I suggest that in cases of evaluative shift concerns that normally point against paternalism – such as a concern with authorship – may direct us in the opposite direction.

A second implication concerns the idea of authorship, or control over one's life. This may seem a stronger point on which anti-paternalism may stick; I surely cannot deny, even in cases of evaluative shift, that paternalism obstructs a person's ability to live as she sees fit.

I certainly do not deny it. But I insist that, just as with happiness, there is a potential symmetry with regard to a person's ability to live as she sees fit in at least some cases of evaluative shift. Consider again Bou-Habib's imagined character, whose youthful decisions leave her very badly off in old age. It is certainly true that if we refuse to intervene, we allow her some authorship over her life. But that authorship occurs entirely and exclusively in her youth. It is not at all true, given her very different priorities, that she can be described as living as she sees fit in her old age, for the simple reason that she lacks the capacity.

### 4.3 Respect

Finally, the third objection suggests that paternalism involves some kind of disrespect for the agent. But it is not clear why we should regard paternalistic intervention in this kind of case as embodying a judgement that the agent is defective in her ability to judge or act, as Shiffrin has it. So if the claim that paternalistic interventions are disrespectful is based on the idea that such interventions necessarily involve judging an agent as defective, then the justification I have offered of paternalism is in fact not disrespectful. My justification has not relied on our insisting that the youth-focused view of the good is *mistaken*, although it might be, or that an agent who makes it is incompetent, although they might be; it is consistent with this policy that, if our agent would maintain a youth-oriented evaluative stance even into her old age, we *should* respect her decisions.

One might think that there is something necessarily disrespectful in interfering with an agent's decisions and actions. But I do not see why that should be the case, if no judgement of incompetence or irrationality is being made. When we intervene with someone's intended behaviour because it will harm *others*, for instance, we do not necessarily conclude that her aims are irrational, or that she is somehow incompetent. We do not even necessarily claim that her aims are immoral. As such, this intervention is not taken to be disrespectful.

Of course, there is a distinction between the two cases. In one case an agent is prevented from acting for someone else's benefit; in another case she is prevented from acting *for her own good*. And an anti-paternalist might insist that this necessarily embodies a form of disrespect. In the case of evaluative shift, however, I am not sure this is true.

Consider a case where paternalism more obviously involves disrespect. One might justify paternalistic interventions in all cases where the balance of benefit was, according to some objective scale, worse than it could be. This is disrespectful because it ignores the agent's own estimation of what weight to place on benefits at different times. In many cases there is only one

such estimation, held constantly throughout the relevant period, including at times when the agent incurs costs. When we intervene paternalistically, then, we are suggesting that the agent's only judgement on the case is flat-out incorrect, and so at least implying that her judgement is faulty

But in the case of evaluative shift, no such judgement is made. It is not that Bou-Habib's youth-oriented person is somehow incapable of understanding the repercussions of her current values, or unable to bring herself to change her ways. And the case for intervention in this case is not that she has made an evaluative mistake. The case for intervention is that there are *two judgements that conflict*. Further, I have not claimed that we should outright ignore the agent's evaluative stance either in the present or the future. Rather, I have said that we should take both stances into account, and that the future stance may often win; these are the cases where paternalism is justified. I do not see how this can be disrespectful. Instead, the argument for intervention rests on respect for her later agency, which involves a very different set of judgements. Indeed, we might think that the agent's ability to judge and act, and her capacity for self-respect (c.f. Waldron) is severely compromised if we allow her to make such significant sacrifices.

One might, of course, still object here that there is still a form of disrespect involved. We still deny an individual the right to live her life as she sees fit, and we still refuse to allow the relevant person full control of her own affairs. However, this seems to rest to some degree on the claim embodied in the previous objection, that people ought to be the authors of their own lives. As I have already argued, the interest people have in being authors of their own lives is not best served by allowing them to entirely bind their future options, even when their evaluative stance will be radically different in the future. And if respecting that capacity is not best served by the strong anti-paternalist stance, moderate paternalism need not involve disrespect.

Still, we might wonder why we ought to consider a person's future preferences at all. An anonymous reviewer suggests the following case: a person has previously wanted an intervention at time  $T$  but, when  $T$  comes, opposes it. Surely we would not even consider intervening out of

respect for the person's earlier evaluative stance unless we had some additional reason to regard that stance as more authoritative (e.g. the present desire is for some reason not autonomous)?

I entirely agree with this evaluation. But there is an important difference between the cases I suggest, and the one just outlined. In the latter, all costs and benefits are felt by the agent in the present. In the former, costs and benefits accrue at both stages. We can try to imagine a case where, even though a person will undergo an evaluative shift in the future, none of the costs and benefits of an intervention will be felt by her once that has occurred. If it is true that there are genuinely no implications for the person in the future (not even, say, psychological implications of shame), then I would agree that we have no reason to intervene. But this seems to me not to be the case in most situations of evaluative shift. It is also worth noting that in the reviewer's case, ignoring the earlier evaluative stance does not imply any kind of agential disrespect. We are not ignoring the stance because we deem it irrational or otherwise defective. We ignore it because it is simply no longer relevant to the question at hand.

#### **4.4 Some further objections**

In the remainder of this section, I consider two more general objections to the argument offered so far, before summarising my claim that there is a *prima facie* case for paternalistic intervention in some cases of evaluative shift.

First, for all I have said thus far, an intervention might come either in the form of paternalistic restrictions in the agent's youth, or rescue in her old age. And one might think that only the first of these responses is properly paternalistic, since the latter respects the agent's ends both when she has her youth-oriented view of the good, and when she does not. So my argument, so far as it goes, might only support rescue rather than paternalism.

This corresponds to an analogous problem considered by Bou-Habib.<sup>4</sup> He notes that in protecting agents' capacities for autonomy, we face a choice between preventing them from acting in ways that will endanger their fundamental capacity for autonomy, and rescuing them from the consequences of those actions. His response is that in many cases, the cost to the rest of us of rescue will be considerably greater; we are thus justified from a self-interested perspective in choosing the option that is less costly to us, so long as both are permissible. The form of this reply seems readily applicable to the case of paternalism. Both paternalistic intervention *and* rescue after the fact are justified by the arguments offered above. All else being equal, it may be preferable to engage in rescue, since this is not paternalistic. But all else is not equal in cases where costly rescue may be required; when two options are significantly different in costs to us, we are justified in preferring the less costly, even if it is somewhat less desirable in other ways. In such cases, paternalistic constraint of the agent's choices is justified.

Second, someone might grant my argument that in many cases of evaluative shift, we must allow some failure of authorship, some diminution of happiness, and some disrespect, at *some* point, either earlier or later. But even if we must choose, they might insist, paternalism is only warranted if we have reason to defer to people's later perspectives. And there is no such reason.

It is certainly true that a stance which claimed to justify paternalism in all cases of evaluative shift would commit such a fallacy. But I also think that a stance which says, in response to the facts of evaluative shift, that the objections I have considered always provide reasons to avoid paternalism makes a symmetrical mistake; it assumes that in a case where someone must suffer a diminution

---

<sup>4</sup> *Supra* note 3, Bou-Habib strongly opposes paternalism, and bases his justification for intervention on the impersonal value of autonomy. This alternative defence is complicated, and would take us too far from the discussion at hand, so I will not try to outline it here. But at least this aspect of his argument seems open to use by a paternalistic view, even if he would reject such usage.

of respect, a loss of welfare, or a loss of control over their life, it should obviously be in the future. And I have no idea why this should be.

Two differences between these stages are that the earlier stage is temporally prior to the later, and that it is the view that the agent holds when the paternalistic intervention would take place. But neither of these features seems morally relevant. The mere fact that one intervention is later than another is irrelevant. And while the fact that the present agent will suffer the intervention does matter, so too does the parallel fact that the future agent will suffer the costs of sacrifice. In such cases, then, it seems to me legitimate to appeal to a balance of interests, and to consider at which stage the agent will incur the greatest loss. I do not suggest that this is always an easy decision to make; but since the view I am aiming to reject suggests that paternalism cannot be justified, at least in part due to the objections outlined in Section 2, it seems to me enough that there will be some cases where their support is either more mixed, or even where it points more strongly in favour of intervention.

#### **4.5 The *prima facie* case for paternalism**

I suggested in the introduction that undermining the force of the three kinds of objection considered in Section 2 provides us with a *prima facie* justification for paternalistic intervention. Before moving onto the question of the scope of this argument, it is worth briefly setting out this case in an explicit manner, summarising the considerations outlined in this section.

In cases of evaluative shift, an agent will suffer in a later stage of her life because of decisions she currently makes, and which she will at that later time repudiate and see as evaluatively mistaken. But at the time when the agent both suffers the relevant harms and holds the evaluative views that would have motivated her to avoid those harms, she is unable to prevent them, since their cause lies in the past. In general, I assume that we have a *prima facie* reason to either relieve or prevent suffering for agents who are unable to do so themselves. Since we have some reason to intervene,

we must choose between a paternalistic intervention before the harm is caused, or a ‘rescuing’ intervention after the fact. Several considerations contribute to preferring paternalistic intervention at least in some cases.

The first reason is simply that paternalism in such cases is less plausibly morally impermissible than it is in cases that do not involve evaluative shift. If paternalism was morally impermissible, we would have to choose between intervening once the harm had been suffered, or not intervening at all; but since three central objections to paternalism do not obviously decide between these two options, it is less plausible to claim that paternalism is impermissible.

Second, I have argued that in some cases, several of these considerations may actually *support* paternalistic intervention. An agent’s well-being and capacity to control the way her life goes may be more fundamentally affected by refusing to intervene than by intervening, while in cases where an agent’s evaluative stance will change radically, it may be epistemically arrogant of *her* to insist that she knows best what will be good for her even on a subjective understanding.

Finally, in a case where we are confronted with two options that have similar levels of moral acceptability, it seems plausible to suggest that we are sometimes permitted to choose the one that is less costly to us. And it will often be the case that a commitment to rescuing someone who is suffering from a decision that they made prior to evaluative shift is significantly costlier than paternalistically preventing that decision in the first place.

## **V. The scope of the argument**

I have outlined an argument that undermines the force of some central objections to paternalistic intervention in some cases. This argument relies on a combination of the fact that an agent will undergo a significant shift in her evaluative stance, and on the fact that rescue would be excessively costly to the rest of us. This section considers the scope of my argument for paternalism. I suggest

a potential extension of the argument from cases of certainty to cases of mere risk, and outline some thoughts on the limits of the argument.

One might think that in most cases, the claim that it is a fact that the agent will undergo an evaluative shift cannot be established. After all, some people surely do go through their entire life, including old age, with a youth-oriented view of the good. My argument as stated thus far does not justify paternalistic intervention in their case. But since we cannot know for sure that any particular individual will undergo an evaluative shift, doesn't this imply that the argument cannot be applied in practice in any case, even if *most* people with such views of the good will undergo an evaluative shift?

I suggested in Section 1 that many paternalistic interventions apply to cases where we share the agent's ends, and agree on their means to that end, but disagree about how much risk is worth taking for a particular end. An action might thus be paternalistic purely because it constrains an action on the grounds of its being 'too risky', where the agent concerned disagrees with that latter assessment. A related response is available to the issue outlined in the previous paragraph. While it is true that we cannot be certain that our youth-oriented agent will undergo an evaluative shift, there is nonetheless a significant risk of that happening. In most cases, then, our paternalistic intervention will be justified not by the knowledge that the agent *will* undergo an evaluative shift, but by the concern that the risk of her undergoing an evaluative shift is too great, given the kind of old age she is setting up for herself. Just as the agent in the case where we can be certain of evaluative shift imposes on herself a life that she will not endorse, and which will be deeply unpleasant, the agent in a case where an evaluative shift is not certain but is significantly risky imposes this *risk* on herself in the future.

Since paternalistic interventions can apply to cases of risk as well as certainty, this amendment does not undermine the status of the intervention as paternalistic. Might it not change the force of the argument, however? To take just one example, I suggested that what undermined the objections

from Section 2 is the fact that a failure to paternalistically intervene would fall foul of the third objection to paternalism, failure to respect an agent's desire to live according to her own view of the good. But if the agent will not in fact undergo an evaluative shift, then our failure to intervene *does not* fail to respect the fundamentally different view of the good life that she has in her old age, since she has no such different view. We might wonder why we should respect merely possible views that an agent might hold, particularly when weighed against her actual, present evaluative stance. Respect is in this sense quite unlike harm; while we have moral reasons to avoid the risk of harm that hold even when no harm will in fact occur, it is less obvious that we have reasons to respect evaluative stances that nobody will actually hold.

I agree that we do not have reason to respect evaluative stances that nobody holds or will hold. But this does not undermine the argument I have offered. When we take into account a future evaluative stance that somebody might hold, we are giving respect not to some hypothetical future person, but to the actual person in front of us. A complete failure to consider ways that somebody's evaluative stance might change is effectively to say that we see their future preferences as morally relevant only if they are consistent with their current evaluative stance. That is not an attitude of respect. I do not mean here to gloss over the fact that there are difficult questions about just how likely an evaluative shift must be to warrant such an attitude. But that is a fundamentally different question than the issue of whether evaluative stances that are merely possible have moral relevance at all. My answer is that they clearly do.

Consider the analogous case with regard to harming others. What justifies our intervening with your action when it will harm others is, let us assume, that they have a right not to be harmed simply because it will benefit you. But in cases where you merely risk harm to others, it will sometimes be that we constrain you even though your act would not in fact have harmed anyone.

In such cases, the justification for our constraint in the case of certain harm to others is either absent or, at best, the principle from which the justification for our constraint in this case is

indirectly derived. And while there are certainly details to work out about what level and kind of risk it is permissible to coercively prevent, this does not seem sufficiently problematic to rule out the idea of coercively preventing people from acting in ways that risk harming others. The same seems true of paternalistic intervention; even if we cannot use the actual failure to respect the agent's ability to live according to her new view of the good in her old age (because she will in fact have no such new view) to directly justify our action, it can still be the principle from which our actual justification is derived; and that actual justification will rest on the idea of risk.

I noted in Section 4 that there will be cases where neither the agent, nor anyone else, has much basis to predict whether she will undergo evaluative shift. In the cases I am now discussing, it should be clear that epistemic considerations alone cannot decide the case. A paternalist may feel, justifiably, that they know more than the agent does about the likelihood of evaluative shift; but whether that risk is sufficient to warrant intervention is not something that can be decided by epistemic considerations alone. It may be that it takes a greater level of future hardship to justify paternalism in a case of mere risk than it does in a case of certainty. Indeed, it may be that (to the extent that we are able to finely assess risk) the severity of risked harm that justifies paternalism varies continuously with the risk of harm. This invites the question of what kinds of harms justify paternalistic intervention when an evaluative shift is certain, probable, or merely possible.

An extreme view is that if an evaluative shift has any possibility of occurring, we may justifiably prevent a person from performing an act that risks any degree of harm, including mild distress. If we have some reason to think that Jane will come to regard her tattoo as crass and offensive, not merely coming to regret it but seeing the considerations that seemed to support it at the time as grossly mistaken, then this view would say that we may paternalistically prevent her from getting the tattoo. This view is too paternalistic. For one thing, even if Jane will definitely undergo an evaluative shift, her tattoo is unlikely to undermine her ability to live her life according to her new view of the good. Even if it is (say, because it is a facial tattoo that excludes her from many

professions), it is unlikely to be of such considerable difficulty to get rid of that Jane could not affordably get rid of it herself.

Moreover, the argument for paternalism I have offered is not a moralistic argument, because it does not claim that the agent *owes* anything to herself in the future. If it were, we might entirely discount the preferences and desires of Jane when she gets her tattoo, or of our youth-oriented agent in her youth, on the grounds that they fail in obligations to themselves in the future. But since this is not the structure of my argument, we ought not to entirely disregard the preferences of the agent at that time. Interventions on such minor issues as an individual's tattoo seem likely to cause significantly more harm than benefit to the agent overall, and in ways that do not avoid significant harm in the future.

Finally, there is a clear difference between this case and the case of the agent with a youth-oriented view of the good, which is that many people who get tattoos when young do not even come to regret them later on, let alone undergo a significant evaluative shift with regard to them. Whereas our youth-oriented agent was merely happy to *accept* the costs associated with her youthful actions, someone with a tattoo may well actively want that tattoo in their later life. So while there is very little risk if we intervene to the agent at a later time in the case of a youth-oriented view, there is considerable risk to the agent at the later time in the case of the tattoo, and other minor decisions.

An alternative view at the more permissive end of the spectrum of possibility could appeal to some kind of sufficientarian principle, saying that we *at least* have reason to paternalistically intervene when there is a considerable risk of someone's current action leading to them later falling below a threshold of sufficiency *and* of them having a significantly different evaluative outlook towards the original action. Such a view would clearly need to offer both an assessment of what level of risk is 'considerable' (see, e.g., Shrader-Frechette 1991; Cranor 1997; Hansson 2004), and of what grounds the sufficientarian threshold, and where it is set (see, e.g., Frankfurt 1987; Huseby 2009; Shields 2012). Both of these questions are complex, and I will not attempt to tackle them here.

But this view seems to me to be a reasonable description of a minimal case where we are permitted, and perhaps even obligated, to intervene on paternalistic grounds.

## **VI. Conclusion**

I have argued that paternalism is justified in some cases where the effects that we deem bad for the agent, but which she deems good for her (at least on balance) at the time of the decision, will occur at a time when she has undergone a significant change in her evaluative stance. Because such cases involve reasons that rely on at least some of the considerations that normally militate against paternalism – i.e. respect for the agent's ability to live life according to her own values; respect for her capacity for agency and self-respect – this case should persuade even those anti-paternalists who are unconvinced by other arguments, such as an appeal to the agent's happiness, or an assumption that agents with certain evaluative outlooks are necessarily incompetent. I offered a suggestion for how we might expand its scope, by an appeal to the normative links between harms and risks of harms. As such, the appeal to evaluative shift provides an argument that justifies paternalism in an important range of cases, and on terms that should appeal to even committed opponents of paternalism.

*My thanks to Jake Wojtowicz and two anonymous reviewers for their feedback on this paper.*

## **Bibliography**

Arneson, R. (2005). 'Joel Feinberg and the justification of hard paternalism?' *Legal Theory* 11: 259–284

Bou-Habib, P. (2011). 'Distributive justice, dignity, and the lifetime view', *Social Theory and Practice* 37: 285-310.

Buchanan, R. T. (2014). 'Right to die: Belgian doctors rule depressed 24-year-old woman has right to end her life', *The Independent*, 2<sup>nd</sup> July 2015.

URL=<http://www.independent.co.uk/news/people/right-to-die-belgian-doctors-rule-depressed-24-year-old-woman-has-right-to-end-her-life-10361492.html>

- Conly, S. (2013) *Against Autonomy: Justifying Coercive Paternalism*. (Cambridge: Cambridge University Press)
- Cranor, C. (1997). 'The Normative Nature of Risk Assessment: Features and Possibilities', *Risk: Health, Safety & Environment* 8: 123–36
- Dworkin, G. (2014). 'Paternalism', *Stanford Encyclopedia of Philosophy*, URL=<http://plato.stanford.edu/entries/paternalism>
- Feinberg, J. (1984). *The Moral Limits of the Criminal Law: Vol 1, Harm to Others* (Oxford: Oxford University Press)
- Frankfurt, H. (1987). 'Equality as a moral ideal', *Ethics* 98: 21–42
- Hansson, S.O. (2004). 'Weighing Risks and Benefits', *Topoi* 23: 145–152.
- Huseby, R. (2009). 'Sufficiency: Restated and defended', *Journal of Political Philosophy* 18, 178-97
- Mill, J.S. (1859). 'On Liberty', in M. Warnock (ed.) (2003). *Utilitarianism and On Liberty* (Oxford: Blackwell Publishing): 88-180.
- Shields, L. (2012). 'The Prospects for Sufficiency', *Utilitas* 24: 101-17.
- Shiffrin, S. (2000). 'Paternalism, Unconscionability Doctrine, & Accommodation', *Philosophy & Public Affairs* 29: 205-51.
- Shrader-Frechette, K. (1991). *Risk and Rationality. Philosophical Foundations for Populist Reforms* (Berkeley: University of California Press).
- Sunstein, C. and Thaler, R. (2008). *Nudge: Improving Decisions About Health, Wealth, and Happiness* (Yale: Yale University Press).
- Tännsjö, T (1999) *Coercive Care: Ethics of Choice in Health & Medicine* (London: Routledge)
- Waldron, J. (2014). 'It's all for your own good', *New York Review of Books*, 9<sup>th</sup> October. URL=<http://www.nybooks.com/articles/archives/2014/oct/09/cass-sunstein-its-all-your-own-good/>.
- Wallace, R. Jay (2013). *The View from Here. On Affirmation, Attachment, and the Limits of Regret* (Oxford: Oxford University Press).