Editorial: Can identity-relative paternalism shift the focus from the principle of autonomy?

In bioethical discourse, JS Mill's proscription that 'the only purpose for which power can be rightfully exercised over any member of a civilised community, against his will, is to prevent harm to others' has become almost axiomatic.[i] Bolstered by the rise of patient autonomy during the mid-twentieth century, Millian conceptions of freedom have become so embedded in bioethical theory, that attempts to justify paternalism have typically involved making one of two claims. Either, they have involved refuting the significance of autonomy as a bioethical principle, and questioning whether it should always be taken to outweigh other bioethical principles. Or, they have sought to cast doubt on the autonomous quality of specific decisions, challenging them as non-autonomous in some important way.

Both approaches are readily apparent in discussions about the moral authority of advance directives, where people have questioned both whether precedent autonomy should always outweigh the person's current best interests,[ii] and whether it is really possible to anticipate in advance the full nature and implications of the decision now at issue. However they have also featured in discussions about contemporaneous decisions, where there has been a tendency to query whether deficits in the autonomous nature of the decision — whether caused by a lack of information, the impact of disorder, or the undue influence of third parties — are sufficient to warrant overriding it.

Within this context, Dominic Wilkinson's theory of 'identity-relative paternalism' is important. Wilkinson draws on Derek Parfit's reductionist view of personal identity to make the claim that we may be justified in overriding a patient's refusal of medical treatment in certain instances, not because their autonomy is in doubt, but because a patient's future self may sometimes be so psychologically disconnected from their present self, that they ought to be regarded as a different person. Breaking down the Millian distinction drawn between harm to self and harm to others, he argues that where a person's medical decision causes harm to a person's future self, this should sometimes be treated as though they are causing harm to another person. And so, Wilkinson concludes, '[i]ndividuals should be prevent from doing to future selves (where there are weakened prudential unity relations between the current and future self) what it would be justified to prevent them from doing to others.'[iii]

'Identity-relative paternalism' is therefore important in seeking to justify paternalism without negating the significance of autonomy, either as a general bioethical principle, or to the individual involved. Rather, the focus is on whether interventions are properly regarded as paternalistic (at least in the way JS Mill understood it), at all. If we do not see the person who will bear the future harm as sufficiently psychologically connected to the person making the choice, then it is not a purely self-regarding choice at all, but one that affects another, future person. This has led some commentators in this edition to question Wilkinson's use of the language of 'paternalism'. Ben Saunders, for example, is doubtful that Wilkinson's argument really renders paternalism permissible. Either, he argues, a person's future self is a 'different person', in which case intervention might properly be regarded as protecting harm to others, and so is not genuine paternalism. Or, the future person is the same person, in which case the interference would be paternalistic, but under Wilkinson's theory, it would not then be justifiable.[iv]

Regardless of the semantics of the theory, analysing contemporaneous refusal cases through the lens of personal identity, rather than autonomy, is significant. As the commentaries in this edition show, however, the extent to which this can enhance the ethical analysis of refusal cases is subject to debate. While Rebecca Dresser praises its contribution, and the guidance it offers to both medical professionals and to the patient themselves;[v] others are more doubtful. Some commentators have

questioned whether Wilkinson is focussing on the most ethically salient dimension of the cases he uses. David Birks, for example, argues that there is an important difference between harm to others and harm to one's future self, namely that harms to others violates 'associative duties' that a person holds towards another by virtue of a valuable relationship, whereas harms to one's future self does not.[vi] Others, meanwhile, have questioned whether the principle of autonomy could do much of the same analytical work here. Saunders, for example, has argued that there is no need to rely on 'controversial metaphysical accounts of personal identity to justify interference'. He suggests that the relevant distinction is not that between harm to self and harm to others, but rather between consensual and non-consensual harms. Freedom is only permissibly restricted to prevent non-consensual harms, however this can include some self-harm where an individual lacks the experience necessary to make an informed choice. Consequently, Saunders argues that Mill's harm principle may already permit some paternalistic interference with some long-term decisions, with the focus on the person's inability to make an 'informed choice' seeming implicitly directed at the autonomous quality of that decision. Similarly, Esther Braun questions whether 'identity-relative paternalism' really enhances the ethical analysis of the examples relied on by Wilkinson, arguing that the practical conclusions that he derives from the theory are already justified by existing ethical principles.[vii] Wilkinson's example of James (whose advanced directive was drawn up on the basis of a dice roll) might, for example, be straightforwardly analysed through the lens of autonomy rather than personal identity.

However while doubts about the autonomous nature of the decision might well explain advance (or 'long-term') decisions, the principle of autonomy is less adept at explaining when (and why) paternalism can be justified in respect of contemporaneous decisions, where the temporal gap between the decision and its consequences is more limited. The examples Wilkinson gives of James (who is refusing the Covid-19 vaccination at great risk to himself) and of Jenny (who wishes for a home birth without medical intervention despite contraindications) are less straightforwardly accounted for by autonomy, or indeed by other bioethical principles. While Braun justifies mandating the Covid-19 vaccination on the basis of potential harm to others, Wilkinson deliberately sets up his example to exclude the possibility of using a harm to others as a 'get out clause.' Likewise in her discussion of Jenny, who is refusing a clinically indicated caesarean section because she wishes for a 'free birth at home', Braun seems to justify intervention through construing it as a case of the doctor refusing to provide treatment that is not clinically indicated. It is not clear, however, what treatment the doctor would be refusing to provide, since Jenny explicitly does not wish for any involvement of health professionals. The key issue is over whether the doctors can override Jenny's refusal to have a caesarean section or not, and Braun is not clear about exactly what ethical principles would permit such an intervention. Yet Jenny's is precisely the kind of case in which the clash between autonomy and paternalism would seem most acute, and navigating the bounds of acceptable intervention, most challenging. It is precisely in cases like this then, that Wilkinson's theory may provide a new way of thinking about the conflict.

Whether 'identity-relative paternalism' is in fact better able to navigate cases is open to question, and it is notable that Wilkinson doubts whether it would justify intervention in the cases of either Jenny or James. Both the commentaries of Eli Garrett Schantz and of Charlotte Garstman and others, raise a number of concerns about the practical reach of 'identity-relative paternalism' and its ability to respond to a range of clinical scenarios. Charlotte Garstman, Sterre de Jong and Justin Bernstein, for example, contend that identity-relative paternalism has 'unacceptably implausible implications'.[viii] Cases involving self-sacrifice (such as the decision to donate one's kidney) could be seen as harming one's future self in a way that, under Wilkinson's theory, could warrant interference, for example; as might a decision to undergo treatment which will enhance one's

quality of time in the short term, but may cause suffering later on. By contrast since a decision to end one's life would not cause harm to a future person, it would be morally impermissible to interfere with one's choice. This leads to the somewhat counter-intuitive conclusion (acknowledged by Wilkinson) that we may be justified in intervening paternalistically where a person wishes to compromise their future quality of life, but not where they wish to end their life entirely. Eli Garrett Schantz, meanwhile, argues that the theory produces unactionable and self-contradictory results when applied to choices where both options present possible harms to one's future self. In these cases, identity-relative paternalism may justify intervening with either decision, and provides no guidance on what to do.[ix] This creates a significant problem in practice, since most important decisions involve options which all pose possible harms to future selves.

While both commentaries recognise various responses that Wilkinson might make to their criticisms, they also identify a common problem, namely that applying the 'identity-relative paternalism' in practice requires us to be able to predict the likelihood of someone's values changing in a way that maintains or disrupts psychological unity, something which it is extremely difficult to do. Schantz, for example, suggests that to overcome the problems identified, the theory may 'index' the justification for intervention to the likelihood and severity of possible harms, and the likelihood of the person's values changing. However predicting how a person's values might change over time is near impossible, as is predicting whether harms may materialise. Consequently, he argues that identity-relative paternalism is 'still unable to render a decision on a case without knowledge of the future which borders on omniscience'. Garstman and others similarly suggest that the most compelling way of addressing their concerns would be to assume that in the case considered, there is simply no reason to suspect that any future person will be a distinct person, since we have no particular reason to believe that there will be a significant break in their psychological unity. Accepting this, however, would seem to significantly narrow the application of identity-relative paternalism. If, in fact, we think most people are likely to have sufficient prudential unity relations between the current and future self, then paternalism can rarely be justified in practice.

This raises important questions about how 'sure' we must be that a person's prudential relations will be weakened before we can act. If the evidential threshold is set too high, then while 'identity-relative paternalism' offers a new way of thinking about cases at a theoretical level, it does little to assist clinicians faced with challenging refusal cases in practice. If it is set too low, it potentially allows quite substantial interference with individual choice, which is hard to reconcile with the weight accorded to autonomy in contemporary bioethics. While Wilkinson recognises this, identifying 'an important question' outstanding 'about whether we should assume prudential unity and give priority to the wishes of the current individual, or assume prudential disunity and prioritise preventing harm to the later self', until this question is satisfactorily answered, identity-relative paternalism's utility as a means of analysing refusal cases will remain a matter of debate.

[i] Mill JS. *On Liberty.* London: Longman, Roberts, and Green, 1864.
[ii] See for example Dresser R. Dworkin on Dementia. Hastings Center Report 1995. 25(6).
[iii] Wilkinson D. The harm principle, personal identity and identity-relative paternalism. *J Med Ethics* 2023. 10.1136/jme-2022-108418.
[iv] Saunders B. Paternalism, with and without identity. *J Med Ethics* 2023. 10.1136/jme-2023-109121.
[v] Dresser R. Medical choices and changing selves. *J Med Ethics* 2023. 10.1136/jme-2023-109120.
[vi] Birks D. Identity-relative paternalism and allowing harm to others. *J Med Ethics* 2023. 10.1136/jme-2023-109118.
[vii] Braun E. Identity-relative paternalism fails to achieve its apparent good. *J Med Ethics* 2023. 10.1136/jme-2023-109119.
[viii] Garstman C. De Jong S. Bernstein J. Does identity-relative paternalism prohibit (future) self-sacrifice? A reply to Wilkinson. *J Med Ethics* 2023. 10.1136/jme-2023-109028.
[ix] Schantz EG. Identity-relative paternalism is internally incoherent. *J Med Ethics* 2023. 10.1136/jme-2023-109009.