It is not about autonomy: realigning the ethical debate on substitute judgement and AI preference predictors in healthcare

Marco Annoni

Interdepartmental Center for Research Ethics and Integrity (CID Ethics), National Research Council, Rome, Italy

Correspondence to
Dr Marco Annoni;
marco.annoni@cnr.it

Received 1 August 2024
Accepted 11 November 2024

ABSTRACT

This article challenges two dominant assumptions in the current ethical debate over the use of algorithmic Personalised Patient Preference Predictors (P4) in substitute judgement for incapacitated patients. First, I question the belief that the autonomy of a patient who no longer has decision-making capacity can be meaningfully respected through a P4-empowered substitute judgement. Second, I critique the assumption that respect for autonomy can be reduced to merely satisfying a patient's individual treatment preferences. Both assumptions, I argue, are problematic: respect for autonomy cannot be equated with simply delivering the 'right' treatments, and expanding the normative scope of agency beyond first-person decisions creates issues for standard clinical decision-making. I suggest, instead, that the development of these algorithmic tools can be justified by achieving other moral goods, such as honouring a patient's unique identity or reducing surrogate decision-makers' burdens. This conclusion, I argue, should reshape the ethical debate around not just the future development and use of P4-like systems, but also on how substitute judgement is currently understood and justified in clinical medicine.

INTRODUCTION

A crucial problem for any model of clinical decision-making based on informed consent is that not all patients are sufficiently autonomous all the time.[1 2]

Patients may be temporarily or permanently incapacitated due to trauma or pathologies, or they may lack the capacity to make sufficiently informed decisions, such as in the case of newborns or individuals with certain mental illnesses and cognitive impairments. In these cases, it may be challenging or impossible to ascertain which treatments these patients would prefer to receive or refuse.

When patients are not sufficiently autonomous to decide for themselves, others must decide on their behalf, assuming the role of surrogate decision-makers—or surrogates.[2] Depending on the clinical context, this delicate role may be filled by a designated person, a caregiver, the patient's family or community, a person legally appointed by a judge, the clinician or the healthcare team.

Surrogate decision-making is a common occurrence in healthcare settings. Estimates suggest that up to 50% of older adults may lack the decision-making capacity at some point in their care, and nearly 70% of treatment decisions near the end of life are made by surrogates.[3 4]

Yet, surrogate decision-making is notoriously problematic. Surrogates may be unable to determine patients' preferences simply because they lack sufficient knowledge and information.[5] Even relatives, partners and caregivers may be unaware of the patient's values and beliefs and thus be unable to decide which therapeutic path would best reflect their preferences and worldview.[6] Moreover, surrogates may themselves have limited cognitive capacities, poor health literacy or insufficient ethical preparedness. Another significant problem is that both surrogates and healthcare professionals often experience high levels of distress when making substitute decisions.[7 8]

In response to these challenges, scholars have proposed new technological solutions, including the development of decisional algorithms to reconstruct the preferences of incapacitated patients.[9]

In 2014, Rid and Wendler were among the first to advocate for a 'Patient Preference Predictor' (PPP), a computerised tool that could infer a patient's treatment preferences by leveraging socio-demographic data.[10 11] This idea was based on empirical evidence showing that certain variables, such as age, gender, income and birthplace, are statistically predictive of healthcare decisions on an aggregate level.[9 11] In essence, individuals facing similar health decisions tend to select options that align with those chosen by others in similar demographic categories.[12]

The proposal to develop a PPP to supplement surrogate decision-making for incapacitated patients has sparked an intense ethical debate, attracting two main criticisms.

The first concerns the PPP's predictive accuracy. Clearly, the case for developing a PPP becomes stronger or weaker depending on whether it is expected to be more or less accurate than traditional surrogates in predicting patients' preferences. Rid, Wendler and colleagues cited a series of empirical studies suggesting that predictions based on demographic variables might be more accurate than those made by next-of-kin.[12–14] Given such preliminary evidence, they argued that it was reasonable to expect a PPP could be at least no worse than standard surrogates in predicting patients' preferences—and potentially better—although further research was warranted.

The second, more substantial criticism, revolves around patient autonomy. A PPP built solely on population-based data risks reducing a patient's moral agency to a set of demographic variables, potentially conflicting with the core principle of respecting patient individual autonomy.[9 15]

To address the limitations, Earp and colleagues recently called for the development of a 'Personalized Patient Preference Predictor' or P4.[9] Unlike the

To cite: Annoni M. *J Med Ethics* Epub ahead of print: [please include Day Month Year]. doi:10.1136/jme-2024-110343

PPP, a P4 leverages recent advancements in machine learning and generative artificial intelligence to extract patients' individual preferences, even those that might be implicit. A P4 would be built using a person-specific corpus of texts, which could include emails, social media posts, blog posts, digital health records and other online activities. These texts would then be used to train a personalised large language model. 'The result would be a kind of 'digital psychological twin' of the person […] that could be queried in real-time as to the patient's most likely preferences for treatment in any given healthcare crisis. In short, the P4 would be a personalized, rather than population-based, patient preference predictor.'[9] According to Earp and colleagues, the P4 would be technically feasible and ethically preferable to the PPP, as it would more accurately reflect individual preferences and sidesteps many of the autonomy-based objections.

While thought-provoking, the proposal to build a P4 to aid surrogate decision-making raises significant practical and theoretical questions. In this article, I contend that the current ethical debate surrounding the justifiability of AI-based preference predictors, like the P4, risks conflating respect for autonomy with other morally significant goods, such as the satisfaction of patient preferences and the honouring of their unique identities. I argue that the assumption that tools like the P4 can respect the autonomy of incapacitated patients by enabling more accurate surrogate judgements is problematic and warrants closer examination.

AUTONOMY, SURROGATE DECISION-MAKING AND SUBSTITUTE JUDGEMENT

Respect for autonomy has consistently been at the forefront of ethical discussions surrounding PPP. Rid and Wendler acknowledged the potential autonomy-related concerns in their original proposal, and much of the subsequent critique has focused on whether these tools respect or undermine autonomy.[10 11 15–17] Earp and colleagues similarly emphasise the autonomy-related benefits of the P4, asserting that it would allow for more accurate surrogate decision-making, thereby respecting autonomy more effectively than the non-personalised PPP.

More specifically, Earp and colleagues argue that a P4 would allow to respect the autonomy of incapacitated patients better than a PPP—and perhaps even better than human surrogates—as it would allow for a more accurate form of 'substitute judgment'.

The substituted judgement standard (SJS) differs from other decisional standards for surrogate decision-making—such as the 'best interest' or the 'rational person' standard—, for it instructs the surrogate to 'make the decision the incompetent person would have made if competent', thereby relying on a hypothetical form of consent.[2 9]

Yet, the SJS introduces a few theoretical challenges, particularly in defining the nature of the autonomy it purportedly respects. How can a P4-empowered substituted judgement respect the autonomy of incapacitated patients, if these patients are no longer autonomous?

To this question, Earp and colleagues reply that 'one way of respecting a person's autonomy in making a substituted judgment is simply to 'get the right answer'—that is, to choose what they would choose in the current situation'. This view aligns with Lamanna and Byrne proposal to take data about incapacitated patients as input 'and derives a confidence estimate for a particular patient's predicted healthcare-related decision as an output', a solution which they term the 'autonomy algorithm'.[18]

The underlying idea is that tools like the P4 or the autonomy algorithm can improve substituted judgements, and thus surrogate decision-making, by providing more precise insights into a patient's individual preferences, thus offering a more accurate way of respecting their autonomy.

For clarity, and to avoid attributing views that others may not necessarily endorse, in what follow I will abstract from the above quotes and focus, instead, on the following argument:

i. Substitute judgement allows to respect the autonomy of incapacitated patients.
ii. Respecting autonomy entails choosing what incapacitated patients would choose.
iii. Choosing what patients would choose requires identifying their treatment preferences.
iv. The P4 enables the identification of patients' treatment preferences.
v. Therefore, using the P4 allows to respect the autonomy of incapacitated patients.

While this argument seems appealing, it is also problematic. Most critiques so far have targeted the ability of P4 models to predict patients' preferences (premise 4). In this article, instead, I focus on the underlying link between respect for autonomy, substitute judgement and individualised treatment preference prediction, questioning two assumptions. First, the view that respect for autonomy can be equated with the satisfaction of individual treatment preferences is questionable whenever such preferences are not actual, explicit and specific (premise 2). Second, that incapacitated patients have a residual autonomy to be respected is not self-evident and requires further justification (premise 1). In the following sections, I argue that both claims require closer examination and call into question the capacity of tools like the P4 to ever respect the autonomy of incapacitated patients.

PRESUMED PREFERENCES AND RESPECT FOR AUTONOMY

At the outset, it is important to clarify why getting patient presumed preferences right is often insufficient to respect their autonomy. Consider the case of an autonomous patient who expresses the preference to receive treatment (x). If the doctor administers treatment (x), she will then satisfy the patient's preferences and respect her autonomy. Conversely, consider the case of a doctor who decides, without informing the patient, to administer treatment (y) because she thinks that it would be more beneficial and cost-effective, despite the patient's explicit consent to (x). In this case, the doctor has not only failed to satisfy the patient's preferences but also disrespected her autonomy. By all account, this is a classic case of medical paternalism. In both examples, it is therefore true that respecting patient autonomy entails respecting their (explicit) treatment preferences.

However, consider the case of a patient who suffers from a vague sense of uneasiness and undefined, fluctuating, symptoms. The doctor is aware that, based on recent evidence, a placebo pill could aid in alleviating these symptoms. She then decides to deceptively prescribe a placebo believing that: (1) would the patient be aware that a placebo could help, she will consent to its use; (2) a placebo is only effective if administered deceptively, and thus deception is the only option to elicit the desired beneficial effects (a belief recently disproved, but that I retain as it is functional to my argument).[19] In this case, the doctor acts on the patient's implicit or presumed preferences, as she does not ask for the patient's consent or explicit preferences. Also, the doctor acts paternalistically, as she openly deceives the patient, infringing on her right to be truthfully informed and non-manipulated. Yet, suppose that (1) is true: the doctor is right, and the patient really would prefer a deceptive placebo

in this situation rather than no prescription. Receiving the placebo would then satisfy the patient's presumed preferences, even though it would disrespect her autonomy. Thus, it is not always true that 'one way of respecting a person's autonomy is to choose what they would choose in the current situation'. One can consistently choose what another person would presumably choose and still fail to respect her autonomy.

These examples show that in discussing the ethics of satisfying patients' treatment preferences two distinctions are key. One is between revealed (ie, explicit) and implicit preferences—or presumed preferences when one's implicit preferences are inferred by surrogates or computer-based tools like a P4 or the autonomy algorithm. The second distinction is between two logically independent goals: respecting autonomy and satisfying the patients' preferences. Often, respecting autonomy involves satisfying patient's revealed preferences, as with informed consent. However, satisfying presumed preferences does not always equate to respecting autonomy, as with the deceptive placebo scenario.

This conclusion extends to substituted judgement via P4. Even granting that a P4 could be accurate in predicting patients' individual presumed treatment preferences, from this fact alone it does not follow that satisfying such preferences would also respect their autonomy.

REVEALED PREFERENCES AND RESPECT FOR AUTONOMY

Before becoming incapacitated patients may have expressed their treatment preferences in various ways, leaving behind written or digital records that a personalised preference predictor might use to infer what they would choose in other scenarios. These predictions would be based on explicit and revealed preferences rather than on presumed or implicit ones. However, even acting on a patient's past revealed preferences may still fall short of respecting their autonomy in any significant sense.

Consider the example of the deceptive placebo once again. Now, imagine the patient had previously expressed her opinion about a similar situation on social media while commenting on a medical drama, stating that 'if I were in the same situation, I would prefer the deceptive placebo'. If the doctor were aware of these posts, would administering the placebo be a way of respecting her autonomy?

Arguably, the answer remains 'No'.[2] Administering the deceptive placebo would still constitute an act of medical paternalism. And while this paternalism may seem more justifiable given the patient's previously expressed preferences, it remains paternalistic nonetheless.[19] This is because respecting autonomy typically requires more than merely knowing and acting on previously expressed treatment preferences; it demands an explicit, first-person, actual and specific consent.

There are several rationales for maintaining this cautionary approach in which respect for autonomy is grounded in actual consent, but two are particularly relevant to the present discussion. One is the well-documented unreliability of previously expressed preferences. This unreliability stems from the inherent difficulty humans face in accurately predicting future states of mind, particularly when their health or circumstances change significantly. Research in decision-making psychology has decidedly shown that individuals frequently misjudge their future feelings, preferences and adaptive capacities.[20] Clinical decisions rarely occur in a vacuum; cognitive, emotional and psychological factors play a significant role, and preferences may evolve in response to contexts. As such, caution is warranted when interpreting previously revealed preferences as reliable.

This concern becomes especially pronounced when relying on data from social media or casual communications, where expressed preferences may not reflect deeply held beliefs and values. For example, someone might declare on social media that they would 'rather be dead than live with a significant spinal cord injury'. Yet, many individuals who later experience such disabilities report a quality of life far exceeding their previous expectations and often choose to live rather than pursue assisted suicide.[20 21] After all, a casual online expression of preference regarding a hypothetical situation is vastly different from a decision in an actual scenario with irreversible consequences.

These remarks are particularly pertinent to the context of preference predictors based on large language models like P4s. These models analyse a patient's corpus of texts—emails, blogs, social media posts and healthcare records—to extrapolate likely treatment preferences. Plausibly, these models would then prioritise past revealed preferences related to scenarios like the one at hand. However, given the inherent unreliability of one's previously expressed preferences in other scenarios, even an accurate prediction extrapolating from such preferences does not guarantee that the outcome would reflect what patients would choose if still competent.[22]

A second rationale to ground respect for autonomy in actual consent is that patients, prior to becoming incapacitated, may have lived a life they never truly endorsed. Individuals trapped in religious cults or abusive relationships exemplify this. Inferring what such individuals would choose for themselves, based on their past digital records, may lead to a misleading prediction. In life-or-death situations, these individuals, if competent, might make decisions that deviate significantly from what they had previously shown to endorse. In such cases, actual consent would offer one possibility of safeguarding their autonomy, whereas substituted judgement based on past revealed preferences would not.

In short, the idea that 'choosing what individual patients would choose' is sufficient to respect their autonomy is inherently suspicious, even if one can extrapolate accurate predictions from past revealed preferences. An autonomous decision expressed in an informed consent does something more than just allowing doctors to choose the 'right treatments' for patients. It is a public act of self-determination, which places ultimate authority over the person impacted by certain choices. As such, it provides an important safeguard against external controlling influencing and provides a very close proxy to one's self-governing will concerning a specific, consequential, clinical decision.

A substituted decision by a third party, even if based on past revealed preferences, is fundamentally different. It may allow to 'choose what one would choose', but it would necessarily lack any first-person agency and specific voluntariness. Yet, in the absence of these elements, it becomes unclear how the respect of one's revealed preferences alone qualifies as respecting their autonomy.

SUBSTITUTE JUDGEMENT, AGENCY AND 'THE ENDORSED LIFE VIEW'

So far, I have argued that the mere satisfaction of a patient's individual treatment preferences (either presumed or revealed) is insufficient to respect patient autonomy. In fact, one can satisfy individual treatment preferences and fail to respect one's autonomy. If this is correct, then it follows that respecting autonomy requires something more than the mere satisfaction of presumed or revealed preferences. For incapacitated patients, however, it is unclear what this 'more' is. Again, how is it possible

to respect the autonomy of non-autonomous patients who have not made any actual, intentional and specific autonomous decision with respect to the choice at stake?

Answering this question depends on which theory, or account, of autonomy one endorses. An obstacle in this respect is that autonomy is a multifaceted concept, and many competing accounts of autonomy exist in medical ethics, as well as in moral and political philosophy.[23 24] In what follows, I depart from Beauchamp and Childress's influential account.[2] This approach is sufficiently shared among defenders and critics of the SJS as an autonomy standard to serve as a plausible common ground for the present analysis.

Beauchamp and Childress conceptualise autonomy primarily as the capacity to make actual, local and sufficiently competent autonomous choices. They distinguish two essential conditions for autonomy: agency (the capacity for competent and intentional action) and liberty (independence from controlling influences). Therefore, in their view, a patient is said to make an autonomous choice if she possesses the capacities to understand the relevant information and implications of her choice (understanding) and to act intentionally (intentionality), free from internal (eg, mental illness) and external (eg, manipulation and coercion) interferences (non-control).[2]

Both critics and proponents of the SJS as an autonomy-based standard agree that it is meaningless to discuss respecting the liberty of incapacitated patients without advance directives (ADs). Phillips and Wendler, for example, assert that 'it is essentially impossible to realize this value in relation to individuals who are no longer capable of making their own decisions and who did not leave any directives indicating how they wish to be treated. While we could follow the principle in these cases, its value is lost'.[25] Similarly, Enoch, in his discussion of the normative force of hypothetical consent, argues that referring to liberty (which he terms 'sovereignty') in the context of unconscious patients 'seems no longer relevant'.[26] Incapacitated patients cannot make an autonomous decision, and thus there cannot be a decision that can be externally manipulated, coerced and interfered with.

Accordingly, proponents of SJS as justified by respect for autonomy tend to conceptualise the autonomy protected by such standard in terms of an expanded notion of agency. One influential proposal in this direction is the 'endorsed life view' approach elaborated by Phillips and Wendler, which instructs 'clinicians and surrogates to make treatment decisions that promote the life patients valued for themselves'.[25] According to this view, although incapacitated patients may have lost the capacity to make their own choices, others can still make 'treatment decisions for them that promote the life they endorsed for themselves'.[25] Thus, while respect for liberty does not apply to incompetent adults, 'respect for their autonomy in this sense does'.[25] Consequently, Phillips and Wendler argue that adopting the endorsed life view allows for respecting the autonomy of incapacitated patients through substituted judgement as a way of protecting their endorsed life view.

Phillips and Wendler's key-argument is that although it is not possible to respect the liberty of incapacitated individuals 'we can still respect their autonomy by ensuring that the course of their lives going forward aligns with their values and beliefs about how their lives should proceed'.[25]

This conclusion can be readily extended to justify the development of a P4 model to supplement substitute judgement from the perspective of respecting autonomy. If we can respect the agency of incapacitated patients in this manner, then a P4 model may offer a novel and potentially more accurate approach to reconstructing their endorsed life views, thereby respecting their autonomy.

Phillips and Wendler's proposal would be compelling if applied only to patients who have previously completed an AD or made prior voluntary decisions about their future care (eg, using or not a P4 to support surrogate decision-making in case of incapacity). Indeed, with respect to these cases in which an actual decision is being made, as noted above, it is generally acknowledged that respect for autonomy entails not only a negative obligation not to interfere with individuals' active control over their actions, but also a positive obligation to respect and foster their agency. Asking in advance whether, and at what conditions, a patient would consent to the use of a P4 in case of her supervened incapacity would indeed be a way of respecting their autonomy 'by ensuring that the course of their lives going forward aligns with their values and beliefs about how their lives should proceed'—that is, would be a way of respecting their agency and thereby their autonomy.

However, this argument is much less persuasive for patients who have left no ADs or clear and publicly recorded treatment indications; and have said nothing about the potential use of tools like the P4 or the autonomy algorithm. In these cases, the endorsed life view seems to suggest that the respect of their agency goes beyond the fostering of their actual autonomous decisions, and instead translates into a positive moral obligation to actively carry on patients' endorsed life plans.

Yet, if respect for autonomy entails a positive obligation to act on other's behalf not only to foster their autonomous decision-making, but to actively carry on their endorsed life plan, this significantly shifts the locus of respect for autonomy to the actions of third parties, expanding the normative relevance of respect for agency while restricting the one of liberty as non-interference.

A problem with this approach is that such an expanded view of agency would then extend to all decisional contexts—not just for incompetent patients. If it is morally significant to ensure that the life trajectories of individuals reflect their endorsed views regarding how their lives should go, even in the absence of their actual decisions, this principle should apply to all agents, including those who are still competent. Otherwise, the autonomy protected by the endorsed life view would be worthy of respect only in cases of incompetence but absent and unworthy of protection for those who retain competence. Clearly, this cannot be the case. Either this form of (weak) autonomy is always worthy of respect, or it is not.

If, however, weak autonomy is always worthy of respect, then a physician in the context of a deceptive placebo scenario could be said to be respecting the 'weak autonomy' of a patient to whom she has deliberately lied. To respect this kind of autonomy, it would only be necessary for the physician, through the deceptive placebo, to ensure that 'the course of the patient's life going forward is consistent with her values and beliefs regarding how her life should proceed'. Instead of a straightforward case of medical paternalism, this would then be framed as a matter of balancing respect for 'weak' (ie, without actual voluntariness) and 'strong' (ie, with actual voluntariness) autonomy. Arguably, such an interpretation would stretch the concept of autonomy too far, blurring important moral distinctions and creating potential confusion in distinguishing between legitimate respect for autonomy and cases of unjustified medical paternalism. By doing so, it would also discount the normative value of informed consent and ADs as specific performative kinds of 'speech acts' meant to safeguard the autonomy of competent patients from external interferences.

To avoid these problematic consequences, there are only two options. One is to provide a different account of the kind of autonomy protected by the SJS. While the above critique of the 'endorsed life view' approach is not a refutation of this general possibility, it at least shifts the burden of proof to those willing to demonstrate that an expanded notion of agency can be coherently defined without leading to unwanted consequences. To my knowledge, this has yet to be shown possible.

The second option, instead, is simply to fall back on a narrower conception of autonomy, one, that is, anchored to actual, local, first-person voluntary decisions, thus renouncing the view that the SJS is primarily a standard justified by the respect of the patient (past) autonomy.

NARROW AUTONOMY, SUBSTITUTE JUDGEMENT AND PERSONAL IDENTITIES

One of the major challenges in adopting a narrow view of autonomy is that, as surrogates, it often seems appropriate to 'decide as the incapacitated patient would have if competent'. This moral intuition is particularly strong when a patient has consistently and vocally expressed specific values and treatment preferences prior to becoming incapacitated but has not left any ADs. In such cases, it feels problematic to disregard deeply held beliefs and endorsed life views, even if the individual is no longer autonomous. The surrogate judgement standard as an autonomy-based criterion provides a prima facie plausible way to make sense of this widely shared intuition, though it introduces the challenge of defining the kind of autonomy that this standard protects.

However, there are other plausible moral grounds for upholding this intuition without invoking autonomy. My limited aim here is not to articulate and defend these alternatives in detail but to demonstrate that there are indeed other moral foundations for acting according to an SJS-like criterion as surrogates, independent of an expanded notion of autonomy.

In this section, I will briefly explore one such alternative: honouring patients' unique identities as a basis for choosing according to an SJS-like criterion. As argued by Broström and Johansson, a primary moral aim in acting on behalf of someone who is no longer autonomous is to honour that person's identity and past agency, essentially to 'show recognition'.[27] This form of recognition is integral to our moral lives, as evidenced by our obligation to respect the values and preferences of those who are deceased. For instance, when selecting a song for a funeral, it often seems important to choose one that the deceased would have selected if they were alive and competent. In this context, it is clear that 'this has nothing to do with the dead having a right to self-governance; it's simply a way of commemorating the person for who they were'.[27] This approach signifies respect for the unique way individuals have engaged with the world and acknowledges the agency they exercised in the past. Thus, when deciding as a person would have, 'we may primarily be making a symbolic gesture aimed at honoring the person and showing respect or recognition for their individuality'.[27]

Honouring one's unique identity, in this sense, differs from respecting autonomy in two significant ways. First, while respect for autonomy emphasises an individual's capacity to make self-governing decisions, honouring identity focuses on recognising one's unique way of existing in the world, which may not always involve the ability for autonomous decision-making. For example, we can honour the identity of children, even though they have never been fully autonomous. In this regard, honouring identity encompasses a broader notion of agency that does not rely solely on the capacity for fully autonomous decision-making. Second, respect for autonomy is inherently future-oriented, whereas honouring identity often has a retrospective focus, acknowledging the kind of person one has been or has become.

Importantly, the view that honouring a person's unique identity is sometimes a moral good worth pursuing in surrogate decision-making is compatible with both a narrow view of autonomy—defined as the local capacity to make actual autonomous decisions—and the moral intuition that surrogates should often choose as the patient would have if they were still competent. In this context, however, the moral purpose of the SJS would not be to respect the patient's autonomy but rather to honour that individual's past identity and agency.

Consequently, developing P4 could provide a novel way to achieve this aim. A P4 could potentially reconstruct a patient's preferences and overall narrative agency in much greater detail than traditional surrogates, especially when the patient has left extensive digital traces and the surrogates are unfamiliar with their past. Thus, developing a P4 could be ethically desirable, even if it is not grounded in respect for autonomy.

Moreover, beyond honouring personal identities, the development of tools like the P4 could also be justified on other grounds, such as respecting the patient's overall best interests based on their past values and beliefs, or reducing the decisional burden and moral distress experienced by surrogates and healthcare professionals, irrespective of the adopted decisional standard.

To reiterate, my claim here is that there are other plausible justifications for the SJS and the development of P4 that do not depend on autonomy. Determining which of these alternatives should be adopted, and how they might be operationalised through new P4-like algorithmic tools based on AI and large language models, is an open question deserving further research. However, such research becomes urgent and meaningful only if we acknowledge that respect for autonomy alone cannot justify the development and use of these tools.

CONCLUSIONS

In this paper, I have critically examined two prevailing assumptions in the current debate over the moral justifiability of algorithmic P4 for substitute judgement in surrogate decision-making. The first assumption is that respect for autonomy can be equated with the mere satisfaction of a patient's individual treatment preferences. The second is the belief that the autonomy of a patient who no longer possesses decision-making capacity—and who has made no autonomous decision regarding the specific choices at hand—can be meaningfully respected through the SJS—with or without the aid of P4-like models.

Both assumptions, I argued, are problematic. Respect for autonomy cannot be reduced to ensuring that the 'right' treatments are provided, and expanding the normative relevance of agency beyond actual, first-person decisions blurs important moral distinctions over cases of medical paternalism.

Consequently, I have advocated for a narrower conception of autonomy, one that is grounded in a patient's actual capacity for autonomous decision-making. This view remains compatible with the idea that a patient's past values, views and treatment preferences should be respected, even when incapacity has supervened. Moreover, it can coexist with the view that 'deciding as the patient would have decided, if competent' is often the moral thing to do as surrogates, especially in instances where such decisions are intended to honour the patient's past unique identity.

In conclusion, while the use of P4 for substitute judgement may not succeed in respecting the autonomy of incapacitated patients, this does not mean that the development of such tools lacks moral or practical significance. On the contrary, P4 systems hold considerable potential to achieve a range of other morally desirable ends. Respecting patient autonomy is not one of them, but this does not diminish the importance of these tools in shaping the future of clinical decision-making.

Acknowledgements I am grateful to the two anonymous reviewers for their insightful comments on an earlier draft of this manuscript.

Contributors I am the sole author of this article. Hence, I am the guarantor. Automatic correction was used to check the grammar and ensure the absence of typos. Additionally, I used ChatGPT to check for residual typos. All the ideas, sentences and arguments are entirely mine.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement No data are available.

ORCID iD
Marco Annoni http://orcid.org/0000-0003-3259-1856

REFERENCES

1 Annoni M. Patient protection and paternalism in psychotherapy. In: Trachsel M, Gaab J, Biller-Andorno N, et al, eds. *Oxford handbook of psychotherapy ethics*. Oxford: Oxford University Press, 2019: 99–110.
2 Beauchamp T, Childress J. *Principles of biomedical ethics*. 8th edn. Oxford: Oxford University Press, 2019.
3 Lepping P, Stanly T, Turner J. Systematic review on the prevalence of lack of capacity in medical and psychiatric settings. *Clin Med (Northfield)* 2015;15:337–43.
4 John S, Rowley J, Bartlett K. Assessing patients decision-making capacity in the hospital setting: A literature review. *Aust J Rural Health* 2020;28:141–8.
5 Shalowitz DI, Garrett-Mayer E, Wendler D. The accuracy of surrogate decision makers: a systematic review. *Arch Intern Med* 2006;166:493–7.
6 Annoni M. Reasons and emotions. In: Boniolo G, Sanchini V, eds. *Ethical counselling and medical decision-making in the era of personalised medicine*. Springer, 2016: 39–48.
7 Su Y, Yuki M, Hirayama K. The experiences and perspectives of family surrogate decision-makers: A systematic review of qualitative studies. *Pat Educ Couns* 2020;103:1070–81.
8 Wendler D, Rid A. Systematic review: the effect on surrogates of making treatment decisions for others. *Ann Intern Med* 2011;154:336–46.
9 Earp BD, Porsdam Mann S, Allen J, *et al*. A Personalized Patient Preference Predictor for Substituted Judgments in Healthcare: Technically Feasible and Ethically Desirable. *Am J Bioeth* 2024;24:13–26.
10 Rid A, Wendler D. Use of a patient preference predictor to help make medical decisions for incapacitated patients. *J Med Philos* 2014;39:104–29.
11 Rid A, Wendler D. Treatment decision making for incapacitated patients: is development and use of a patient preference predictor feasible? *J Med Philos* 2014;39:130–52.
12 Shalowitz DI, Garrett-Mayer E, Wendler D. How should treatment decisions be made for incapacitated patients, and why? *PLoS Med* 2007;4:e35.
13 Smucker WD, Houts RM, Danks JH, *et al*. Modal preferences predict elderly patients' life-sustaining treatment choices as well as patients' chosen surrogates do. *Med Decis Making* 2000;20:271–80.
14 Houts RM, Smucker WD, Jacobson JA, *et al*. Predicting elderly outpatients' life-sustaining treatment preferences over time: the majority rules. *Med Decis Making* 2002;22:39–52.
15 Jardas E, Wasserman D, Wendler D. Autonomy-based criticisms of the patient preference predictor. *JME* 2022;48:304–10.
16 John S. Patient preference predictors, apt categorization, and respect for autonomy. *J Med Philos* 2014;39:169–77.
17 Sharadin NP. Patient preference predictors and the problem of naked statistical evidence. *J Med Ethics* 2018;44:857–62.
18 Lamanna C, Byrne L. Should Artificial Intelligence Augment Medical Decision Making? The Case for an Autonomy Algorithm. *AMA J Ethics* 2018;20:E902–910.
19 Annoni M. The Ethics of Placebo Effects in Clinical Practice and Research. *Int Rev Neurobiol* 2018;139:463–84.
20 Blumenthal-Barby J, Fletcher FE, Taylor L, *et al*. Ethical Complexities in Utilizing Artificial Intelligence for Surrogate Decision Making. *Am J Bioeth* 2024;24:1–2.
21 Blumenthal-Barby J. *Good ethics and bad choice: the relevance of behavioral economics for medical ethics*. MIT Press, 2021.
22 Starke G, Jox RJ. Potentially Perilous Preference Parrots: Why Digital Twins Do Not Respect Patient Autonomy. *Am J Bioeth* 2024;24:43–5.
23 Dworkin R. *The theory and practice of autonomy*. New York: Cambridge University Press, 1992.
24 Christman J. Autonomy in moral and political philosophy. In: Zalta E, Nodelman U, Allen C, et al, eds. *Stanford encyclopedia of philosophy*. Stanford University, 2020.
25 Phillips J, Wendler D. Clarifying substituted judgement: the endorsed life approach. *J Med Ethics* 2015;41:723–30.
26 Enoch D. Hypothetical Consent and the Value(s) of Autonomy. *Ethics* 2017;128:6–36.
27 Johansson M. Is hypothetical consent a substitute for actual consent? In: Rønnow-Rasmussen T, Petersson B, Josefsson J, et al, eds. *Hommage à Wlodek. Philosophical Papers Dedicated to Wlodek Rabinowicz*. 2007.