Received: 15 November 2020 | Revised: 8 February 2021 | Accepted: 28 March 2021

DOI: 10.1111/bioe.12886

SPECIAL ISSUE: PROMISES AND CHALLENGES OF MEDICAL AI

WILEY

Exploring the phenomenon and ethical issues of AI paternalism in health apps

Michael Kühler

Karlsruhe Institute of Technology (KIT), Academy for Responsible Research, Teaching, and Innovation (ARRTI), Karlsruhe, Germany

Correspondence
Michael Kühler, Karlsruhe Institute of Technology (KIT), Academy for Responsible Research, Teaching, and Innovation (ARRTI), Karlsstraße 11, Karlsruhe 76133, Germany.
Email: michael.kuehler@kit.edu

Abstract

Health apps, including consumer-oriented fitness apps, have two functions. They are supposed to monitor and promote users' health, the latter by way of being an instance of persuasive technology. The use of artificial intelligence (AI) allows for AI health apps, i.e., health apps that act more and more autonomously when it comes to analyzing users' health data and arriving at tailor-made results on how to improve their health. Consequently, AI health apps seem to gain a paternalistic potential. This is a game-changer, for corresponding issues of paternalism can then no longer be traced back to human engineers. Instead, the paternalizing party just is the AI system. Hence, AI health apps lead to the novel issue of *AI paternalism* in health care. In this paper, I explore this novel phenomenon and its ethical implications. Firstly, I discuss from a critical perspective whether the notion of AI paternalism makes (conceptual) sense to begin with. Unsurprisingly, I argue that it does and how so. Secondly, I briefly indicate important ethical issues that AI paternalism in health apps raise and which need to be discussed in more detail in order to judge under which conditions (certain forms of) AI paternalism might be considered acceptable, if at all.

KEYWORDS

AI, autonomy, health apps, paternalism

1 | INTRODUCTION

Health apps, including consumer-oriented fitness apps, typically have two functions. Firstly, they track health-related data, like weight, calories, heart rate, steps, etc. This allows users to gain an overview of their health and fitness in terms of such quantifiable criteria. Notably, they allow users to check their personal development over time. For example, users may witness that they are losing weight while regularly exercising, eating healthier, and consuming fewer calories, thus getting quantifiable "proof" of their improving health. In fact, such progress is precisely one of the main reasons to use health apps in the first place. Accordingly, the second function of health apps is to initiate and foster a change in behavior in their users toward a healthier lifestyle. Health apps are, in this regard, an instance of persuasive technology[1] and often rely on strategies of positive feedback, gamification, or sharing results on social media for joint experiences (and allowing for social pressure) to motivate users

[1]See IJsselsteijn, W., de Kort, Y., Midden, C., Eggen, B., & van den Hoven, E. (Eds.). (2006). *Persuasive technology: First international conference on persuasive technology for human well-being, PERSUASIVE 2006, Eindhoven, The Netherlands, May 18–19, 2006, Proceedings*. Springer-Verlag. https://doi.org/10.1007/11755494

accordingly. In short, the purpose of health apps is to monitor and promote users' health.[2]

So far, health apps are for the most part rigidly designed by human engineers who implement a specific set of health-related goals or values—usually to choose among by the user—and a certain way of how to change the user's behavior, e.g., by specific gamification routines or by letting the app make predefined suggestions on exercises or nutritional behavior. However, the ever-increasing use of artificial intelligence (AI) in society,[3] including the health care domain, introduces the possibility of AI health apps, i.e., health apps that include machine or deep learning algorithms, and that may act more and more autonomously when it comes to analyzing users' health data and arriving at tailor-made results on how to improve their health. A first example of this might be Amazon's recently announced *Halo*.[4] In general, such AI health apps would be capable of:

1. *analyzing* their users' behavior in light of the individual user's tracked health data against a more encompassing database of general (quantified) information on human health,
2. "*drawing conclusions*," i.e., computing results, as to which behavior would benefit the individual user (best) against a predefined (quantified) notion of health as well as general information on sufficiently similar persons, and,
3. based on tracked user data, both health and non-health related, and general information on success rates of motivation strategies, *influencing* the individual user's decision-making and behavior accordingly, for example by way of making tailor-made "nudging" suggestions on what to do or by means of gamification.

Arguably, such AI health apps may now be considered sufficiently autonomous actors when it comes to influencing the users' behavior for their own good. If so, it seems that AI health apps gain a paternalistic potential. While it is often claimed that health apps promote users' autonomy in that users have to choose the goals themselves and the app merely functions as a tool to realize these goals more efficiently and successfully, AI health apps are game-changing. When assuming autonomously acting AI health apps, the scenario not only raises the question to what extent such apps may still be thought of as promoting, or at least not undermining, users' autonomy when influencing them for their own good, but it also leads to the problem that corresponding issues of paternalism can no longer be traced back to human engineers. Instead, the paternalizing party just is the AI system.[5] Hence, AI health apps, and to a certain degree AI-supported decision-making in general, arguably lead to the novel issue of *AI paternalism*, particularly in the health care domain.[6]

Surprisingly, current debates in medical ethics and in ethics of technology are still lacking when it comes to acknowledging this paternalistic dimension of AI technology. For instance, it is telling that a paternalistic dimension is not even mentioned in the Nuffield Council's roadmap for future research in AI.[7] In the following, I will explore this novel notion of AI paternalism. Firstly, I will discuss from a critical perspective whether the notion makes (conceptual) sense to begin with, i.e., whether the traditional concept of paternalism can reasonably be applied. Unsurprisingly, I will argue that it can, albeit with some slight modifications. Secondly, I will briefly indicate important ethical issues that AI paternalism in health apps raise and which need to be discussed in more detail in order to judge under which conditions (certain forms of) AI paternalism might be considered acceptable, if at all.

[2]This description of health apps is admittedly mostly positive in tone, as it focuses on the desired positive effects in improving users' health. However, this is not supposed to imply that there is no room for criticism. In fact, the idea of a "quantified self," including its reductionist notion of health, has quite correctly faced substantial criticism concerning our self-understanding as persons. The techniques of gamification and the possibly resulting social pressure when sharing results on social media raise serious concerns about the users' autonomy. Finally, sharing one's data "in the cloud" and letting companies that develop such apps have access to one's personal (health) data raises equally serious concerns about users' privacy. See Sharon, T. (2017). Self-tracking for health and the quantified self: re-articulating autonomy, solidarity, and authenticity in an age of personalized healthcare. *Philosophy and Technology*,*30*(1), 93–121; Danaher, J., Nyholm, S., & Earp, B. D. (2018). The quantified relationship. *American Journal of Bioethics*, *18*(2), 3–19. https://doi.org/10.1080/15265161.2017.1409823

[3]For overviews of the debate about AI's ethical implications, see Sullins, J. (2019). Information technology and moral values. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy (summer 2019)*. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2019/entries/it-moral-values/; Müller, V. C. (2020). Ethics of artificial intelligence and robotics. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy (winter 2020)*. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2020/entries/ethics-ai/

[4]Amazon. (2020). Amazon Halo - YouTube. https://www.youtube.com/channel/UCkGJiALRRLVVIDIQ20RYoHg. Accessed September 7, 2020.

[5]An anonymous reviewer suggested that, even in such cases, the paternalizing party could rather be identified as, for instance, the health insurance company if patients are compelled by them to use an AI health app. However, to my mind, AI health apps would still be game changing if they can be considered sufficiently autonomous actors as described above. A fitting analogy in this regard might be parents who instruct their older child to watch over the younger sibling. Even if the parents might ultimately be considered *a* paternalizing party, delegating the task with its concrete paternalistic interferences to the older child surely makes this child a (sufficiently autonomous) paternalizing party as well. Moreover, imagine cases in which there are no other actors, like health insurance companies, involved, but the AI health app is simply part of any user's default set-up. Both scenarios should, therefore, make clear that AI health apps, indeed, present us with a novel situation, with them being the paternalizing party, after all.

[6]In a way, this might jeopardize the major shift in medical ethics and practice from medical paternalism in the doctor-patient relationship to respecting patients' autonomy, the latter of which is expressed in the requirement of getting patients' informed consent. See Beauchamp, T. L., & Childress, J. F. (2019). *Principles of biomedical ethics* (8th ed.). Oxford University Press, ch. 4.

[7]Cp. Whittlestone, J., Nyrup, R., Alexandrova, A., Dihal, K., & Cave, S. (2019). *Ethical and societal implications of algorithms, data, and artificial intelligence_ a roadmap for research*. Nuffield Foundation. A notable exception is Wagner, N.-F. (2019). Doing away with the agential bias: Agency and patiency in health monitoring applications. *Philosophy & Technology*,*32*(1), 135–154. https://doi.org/10.1007/s13347-018-0313-7, who mentions and addresses a paternalistic dimension. Moreover, under the guise of *epistemic paternalism*, AI-supported decision-making is critically discussed in medical diagnosis and decision-making. See, for example, Grill, K., & Hansson, S. O. (2005). Epistemic paternalism in public health. *Journal of Medical Ethics*, *31*(11), 648–653. https://doi.org/10.1136/jme.2004.010850; McDougall, R. J. (2019). Computer knows best? The need for value-flexibility in medical AI. *Journal of Medical Ethics*, *45*(3), 156–160. https://doi.org/10.1136/medethics-2018-105118; Nucci, E. D. (2019). Should we be afraid of medical AI? *Journal of Medical Ethics*, *45*(8), 556–558. https://doi.org/10.1136/medethics-2018-105281; Bjerring, J. C., & Busch, J. (2020). Artificial intelligence and patient-centered decision-making. *Philosophy & Technology*. https://doi.org/10.1007/s13347-019-00391-6; Axtell, G., & Bernal, A. (Eds.). (2020). *Epistemic paternalism. Conceptions, justifications and implications*. Rowman & Littlefield Publishers.

2 | DOES THE NOTION OF AI PATERNALISM MAKE SENSE TO BEGIN WITH?

Consider the following traditional definition of paternalism:

> "I suggest the following conditions as an analysis of *X acts paternalistically towards Y by doing (omitting) Z*:
>
> 1. *Z* (or its omission) interferes with the liberty or autonomy of *Y*.
> 2. *X* does so without the consent of *Y*.
> 3. *X* does so only because *X* believes *Z* will improve the welfare of *Y* (where this includes preventing his welfare from diminishing), or in some way promote the interests, values, or good of *Y*."[8]

Would the notion of *AI paternalism* fit this concept of paternalism or at least a slightly revised version of it? Based on the definition Dworkin gives, four conditions can be derived that any occurrence of paternalism must meet. Consequently, if the notion of AI paternalism in health apps is to make (conceptual) sense within the traditional definition of paternalism, AI health apps need to:

1. interact with users (Y) intentionally,
2. include a notion of what is (supposedly) good for Y,
3. interfere with the liberty or autonomy of Y, and
4. do so without Y's consent.

Condition 1 already seems to exclude AI systems as possible paternalistic actors, for AI systems obviously do *not* have intentions. Only persons have intentions, i.e., mental states that include a representation of a desired result of one's future action and motivate one to act accordingly.[9] However, certain AI systems, like AI health apps, may certainly be described as showing *goal-oriented* behavior.

As described above, the goal of AI health apps is to promote their users' health. Moreover, AI health apps are defined as being sufficiently autonomous in their behavior, notably in choosing the best (or most efficient) means to reach their goals.[10] If so, a slight revision of paternalism's traditional definition to include goal-oriented behavior suffices to allow the first condition to be met.

Concerning condition 2, AI systems likewise obviously do *not* (consciously) have a concept of the good or uphold any ideas of what is good for Y, not even in terms of health. However, following condition 1, they *do* include a set of computable (quantifiable) criteria that serve as a definition of their goals. This may be understood in terms of a predefined pattern (of development) against which recognized patterns in tracked data can be compared. A match of patterns or an increasing resemblance of the recognized pattern to the predefined pattern may be considered in terms of the AI's (increasing) success at reaching its goal. If so, it may be said that AI health apps behave *as if* they had a notion of what is good for Y. Yet, assuming that the pattern to be achieved is predefined by the AI system's programmers, the (quantified) notion of health—or of the good in general—must be considered an external input and, thus, not autonomously chosen by the AI system itself. In any case, a slight revision to allow for "*as if*" means the second condition is met as well.

When considering condition 3, one might think at first that AI health apps are surely not capable of interfering with their users' liberty, in the sense of freedom of action. After all, how could an app physically hinder one to do what one wants? However, there are numerous apps that are not limited to creating outputs on smartphones or tablets. Consider smart home apps, the purpose of which is to exert physical control over one's home environment. Users may even delegate a certain amount of physical control to them by automating certain aspects, like turning on the lights and music in the morning in order to wake up more easily. Moreover, imagine AI systems in cars capable of recognizing if the driver, upon entering the car, is too tired or drunk to drive safely and then prevent the car from starting. Future scenarios might even include more encompassing possibilities of personal AI assistants interfering with their users' liberties in their daily life. Hence, if linked to other systems capable of certain environmental control, AI health apps could interfere with users' freedom of action, after all.

In addition, there can hardly be any doubt that AI health apps are able to influence their users' behavior. Being an instance of persuasive technology, this is precisely what they are designed for in the first place. They may influence the users' will-formation or decision-making process by making especially enticing suggestions or by raising the users' awareness of certain possible options, while at the same time excluding other options in their suggestions. This holds even more if users are not aware that they are dealing with AI technology. Prevalent current examples are search results or suggestions on Google or Amazon, which are filtered based on algorithms. Accordingly, if users are not aware of the algorithmic filter, they might not even think of critically reflecting upon the results or looking elsewhere as well. And even if users are aware of it, trusting or "blindly" following the results is just so easy and convenient. Hence, users' will-formation and decision-making process are substantially influenced, which has already sparked a lively debate on whether this undermines users' autonomy.[11] Arguably, this holds all the more

[8]Dworkin, G. (2020). Paternalism. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy (fall 2020)*. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2020/entries/paternalism/, sec. 2.

[9]Of course, defining "intention" in detail is much more complicated. For an overview, see Setiya, K. (2018). Intention. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy (fall 2018)*. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2018/entries/intention/

[10]As such, they can be regarded as an instance of autonomously acting technologies, like self-driving cars or so-called "killer robots," which raise troubling questions about the attribution of responsibility. The apparent exclusion of responsible human persons, in turn, s the (rudimentary) agency of AI health apps and their possible role as the paternalizing party. See Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology,6*(3), 175–183. https://doi.org/10.1007/s10676-004-3422-1; Sparrow, R. (2007). Killer robots. *Journal of Applied Philosophy, 24*(1), 62–77. https://doi.org/10.1111/j.1468-5930.2007.00346.x

[11]For a general overview of the ethics of manipulation, see Noggle, R. (2020). The ethics of manipulation. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy (summer 2020)*. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2020/entries/ethics-manipulation/. For an overview of the debate on AI technology manipulating behavior and influencing autonomy, see Müller, op. cit. note 3, sec. 2.2.

if the AI technology in question is supposed to promote users' well-being and users are aware of this, like in the case of AI health apps. Imagine the AI health app making suggestions on what to eat, including corresponding shopping lists and recipes, all based on an analysis of the particular user's health and which eating habits would likely improve it. These suggestions might even acknowledge probable food intolerances, of which the user is still unaware, and thus exclude corresponding ingredients right from the start. If users are confident that trusting or even "blindly" following their AI health app's suggestions will be beneficial for their health, precisely because this is what the app is designed for, this might even be characterized as users delegating the respective choices, and thus their respective autonomy, to the app. Overall, there can hardly be any doubt that the third condition of applying the traditional concept of paternalism to the phenomenon of AI paternalism is met.

The fourth condition again raises conceptual doubts. After all, if users need to install and use an AI health app *intentionally*, this obviously means that they—even explicitly—consent to it. However, it can be argued that a *general* consent to use an app does not necessarily mean that one consents to each and every *specific* influence on autonomy, especially if one is not aware of all specific influences on one's will-formation or decision-making process, like notably the exclusion of specific results and options. As discussed before with regard to the third condition, the app's influence on users' autonomy might be subtler and more complex than a general consent to use the app may cover. Consider again the example of the AI health app excluding certain types of food in its suggestions on what to eat based on its ongoing analysis of the user's health data. It is far from clear that by intentionally using the app in general, the user has also explicitly consented to all such specific exclusions. This is even more obvious if users are not even aware that the AI functionality might comprise such exclusions at some point based on its learning algorithm. After all, how can we consent to something if we do not even know about it?

As an analogy, imagine generally allowing a friend to make suggestions and influence your decision-making in your everyday life. While such a general consent may very well cover that the friend will take action even if not specifically asked to, there is arguably still room for specific criticism if the friend, on specific occasions, purposefully tries to influence you without informing you about it or tries to hide in principle valuable options from you just because he or she thinks that these options will do you no good. If this analogy and the intuition about possible criticism in specific cases are plausible, one may conclude that the fourth condition is met as well, at least when it comes to paternalistic influences in specific cases and notably if users are not even aware of them.

In addition, one might think of future scenarios in which personal AI assistants are ubiquitous and also include the functionality of AI health apps by default, maybe even without this being advertised explicitly or only mentioned in the fine print—which usually nobody reads. Arguably, this puts into question the idea of users having consented to *all* its influences on one's life even more—and such future scenarios might be closer than we realize.

If the above considerations are plausible, all four conditions are met, albeit partially in a slightly revised version. Hence, the notion of AI paternalism, indeed, makes (conceptual) sense. Reformulating and adapting Dworkin's traditional definition of paternalism, AI paternalism in health apps can thus be characterized as follows: *AI health apps are capable of showing a paternalistic goal-oriented behavior toward users in that they (a) might interfere with their liberty or influence their autonomy, (b) of which users might not even be aware or at least have not consented to on specific occasions, and (c) do so because of their goal-oriented programming in terms of promoting users' (quantifiable) health.*

However, while this introduces the general conceptual possibility of AI paternalism, it is helpful to discuss the possible types of paternalistic interference by AI health apps in more detail.[12] In the following, I will discuss two important distinctions in this regard, namely between *hard* and *soft* paternalism, including the notion of *nudge* paternalism, on the one hand, and between *strong* and *weak* paternalism, on the other. The most prominent distinction is likely between *hard* and *soft* paternalism. While hard paternalism marks an interference with a person's liberty, i.e., freedom of action, regardless of whether the person has made an autonomous decision, soft paternalism is intended to check whether a person's decision is (sufficiently) autonomous or to promote the person's autonomy by providing relevant information or insights. Yet, if the person is deciding or acting (sufficiently) autonomously, no further interference or influence is allowed.[13]

When it comes to AI paternalism, both hard and soft paternalism seem problematic to apply. Hard paternalism may only come into play for AI health apps if these are linked to suitable environmental controls, for only then are such apps able to interfere with users' freedom of action. Moreover, when considering soft paternalism, it is hard to imagine how AI health apps would be able to check users' (sufficient) autonomy. However, AI health apps—just as it is argued for "traditional" health apps—may very well be considered to be promoting users' autonomy by providing important information and "insights," namely the results of the individual user's health data analysis, e.g., the likelihood of a specific food intolerance. If so, AI health apps may act in a soft paternalistic way by influencing the user to make more autonomous decisions about, for instance, what to eat.

The latter aspect of influencing users' decision-making leads to another prominent type of paternalism, namely *nudge* paternalism.[14] The main idea of *nudging* persons to make better decisions concerning their health and well-being is to make these better options more attractive or easily accessible, without explicitly hindering persons to act otherwise. This way, people's decision-making process is—more or less subtly—influenced for their own good.

[12]For the following distinctions, cp. Dworkin, op. cit. note 8, sec. 2.

[13]The classical example of a person wanting to cross a bridge, not knowing that it is unsafe, which is why we are allowed to hinder her at first and inform her about the bridge's unsafety, can be found in Mill, J. S. (1859). On liberty. In J. M. Robsen (Ed.), *Thecollected works of John Stuart Mill, volume XVIII – Essays on politics and society part I* (pp. 213–310). University of Toronto Press, 1977, p. 294.

[14]See Thaler, R. H., & Sunstein, C. R. (2009). *Nudge. Improving decisions about health, wealth and happiness* (revised and expanded edition). Penguin.

As an instance of persuasive technology, AI health apps may certainly be regarded as a prime example of such nudging paternalistic influence. As mentioned above, AI health apps may, for example, make a specific suggestion on what to eat, show the respective recipe, and add a corresponding shopping list. This undoubtedly makes it very convenient for users simply to follow this suggestion and, thus, opt for the better option in terms of their health.

The second important distinction is between *strong* and *weak* paternalism. While weak paternalism is defined by only interfering with the means people employ to pursue their goals, whatever these may be, strong paternalism also puts into question and interferes with what people pursue as ends. For example, my friend might interfere with my decision to take the bus to be at a concert in time because she thinks the bus will take too long, which is why she orders me a taxi. This is a case of weak paternalism because the goal of attending the concert at all is not questioned. However, if she prevented me from attending the concert because she thinks it will be a waste of my time and I should be doing something more meaningful instead, this would be an instance of strong paternalism.

AI health apps would apparently show a strong paternalism, as health is the only goal and users are precisely influenced to pursue it (better). However, this does not imply that health is treated as an end. It may also be regarded as a means only, even if it may be considered an all-purpose means that everyone should pursue for instrumental reasons. In any case, AI health apps are simply incapable of distinguishing between health as a means and health as an end. Their goal-oriented behavior does not include such a specification, even if only because there are no other (preprogrammed) goals that could force a comparison. Consequently, AI health apps' paternalistic influence cannot meaningfully be described as either strong or weak paternalism, although users might be tempted to reflect on the value of health in their lives because of it, i.e., whether they consider health in their lives to be an (all-purpose) means or an end to be pursued for its own sake. Moreover, because of this inability of AI health apps to specify health as a means or an end, there is also no possibility of a combined strong and soft paternalism because this would imply a paternalistic influence that aims at a critical reflection on the (quantified) notion of health within the broader context of users' well-being and good life, i.e., to address the normative question of which aspects should be pursued in life as ends. The (quantified) notion of health remains a preprogrammed, i.e., externally predefined goal of the AI health apps' goal-oriented behavior.

3 | ETHICAL ISSUE OF AI PATERNALISM IN HEALTH APPS

The latter point leads to the first important ethical issue of AI health apps' paternalistic influence. An externally defined notion of what is good for a person has traditionally been the crucial point of criticism of any form of paternalism—at least if the person interfered with is a sufficiently autonomous adult.[15] Respecting someone's freedom and autonomy means letting persons make their own decisions in life as long as no one else is negatively affected.[16] However, it should be noted that still not all forms of paternalistic interference are shunned in modern medical ethics. For instance, the requirement of getting patients' *informed* consent allows for soft medical paternalism, as patients' initial wishes are not just taken and accepted at face value. Instead, patients need to be adequately informed about treatment options and their practical implications, so they can make a sufficiently informed and thus *more autonomous* decision. In extreme cases, even hard paternalistic interferences are defended.[17] Still, modern medical ethics and a liberal point of view more generally certainly put the onus on paternalism.

Consequently, the (quantified) notion of health in AI health apps is in need of a strong enough ethical justification in order to be acceptable. Yet, AI health apps cannot provide such a justification themselves, even if they may be considered sufficiently autonomous actors and capable of paternalistic influence. This is, firstly, due to agential constraints. AI health apps are not full-fledged persons capable of engaging in a meaningful debate about normative reasons. Secondly, and relatedly, the (quantified) notion of health needs to be externally defined and preprogrammed, which essentially calls for a normative social debate about how to define health.[18] Even if there is sufficient room for a normative social consensus on a broad definition of health and its importance in people's lives, there will arguably always be conflicting positions concerning, at the very least, some of the details. A responsible design of AI health apps, therefore, needs to pay close attention to this debate about the notion of health and especially its (reductionist) quantifiability in order to render AI health apps' paternalistic influences acceptable.

Additionally, even if health is taken as an important value—if only in terms of a valuable all-purpose means for everyone—it is certainly not the only ingredient of people's well-being and of leading a good life. The normative social debate on health must, therefore, also include a critical reflection on health's importance in comparison to other personal values and in light of more encompassing accounts of well-being and how to live a good and meaningful life.[19] Clearly, this makes matters even more complex, but a responsible design of AI health apps needs to take this more complex critical reflection into account as well if a problematic reductionism of possible and

[15]See Kant, I. (1793). On the common saying: That may be correct in theory, but it is of no use in practice. In M. J. Gregor (Trans.), *Practical philosophy* (pp. 271–309). Cambridge University Press, 1996, p. 291; Mill, op. cit. note 13, pp. 213–310. PP. 223f.

[16]This is essentially Mill's *harm principle*. See again Mill, op. cit. note 13, pp. 213–310. PP. 223f.

[17]See influentially Beauchamp & Childress, op. cit. note 6, ch. 4 & 6. For an even more striking defense of hard paternalism, see Conly, S. (2013). *Against autonomy. Justifying coercive paternalism*. Cambridge University Press.

[18]See Murphy, D. (2020). Concepts of disease and health. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy (summer 2020)*. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2020/entries/health-disease/

[19]See Crisp, R. (2017). Well-being. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy (fall 2017)*. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2017/entries/well-being/; Fletcher, G. (Ed.). (2017). *The Routledge handbook of philosophy of well-being*. Routledge.

legitimate personal values of AI health apps' paternalistic influence is to be avoided.

However, any attempt at doing so likely faces the problem of AI technology's *opacity* in terms of how results are reached.[20] While this may be addressed in making the (quantified) notion of health, used as the preprogrammed goal of AI health apps, as well as its justification as transparent as possible, the AI health apps' derivation of specific paternalistic influences on particular users' autonomy will likely remain opaque. This is, of course, a general problem when employing AI technology, but AI health apps' paternalistic influence makes it all the more pertinent.

Moreover, AI's opacity leads to another ethical issue when considering AI health apps' potential for soft paternalism, understood in terms of aiming at the promotion of users' autonomy. If the AI's resulting suggestions and influences on improving users' health are not or even cannot be translated into meaningful reasons and insights, i.e., ones that are understandable in terms of human reasoning, AI health apps' potential for soft paternalism is considerably limited and the latter's possible justification is likely jeopardized. Hence, a responsible design of AI health apps needs to aim at some form of meaningfully translating the AI system's results into humanly accessible reasons. Arguably, it might already be helpful—if possible—to make recognized health patterns, like food intolerances, explicit, based on defining corresponding patterns as such. Otherwise, the use of and especially "blind" trust in AI health apps' paternalistic influence might lead to an *infantilization* of users and, instead of a soft paternalistic promotion of users' autonomy, actually undermine and diminish it.[21]

Furthermore, such "blind" trust in AI health apps' accuracy might be misguided anyway, once one acknowledges the possibility of *algorithmic biases*.[22] This is especially problematic if such biases remain unnoticed. The (quantified) notion of health might very well be inappropriate for certain groups of people, and even if there are different definitions for different groups of people, the AI's application to individual users might still be off at least to some degree due to the more or less encompassing and fitting data to which it has access. If so, corresponding paternalistic influences might *not* be beneficial for particular users, after all. Consequently, a responsible design of AI health apps must include making such possible biases explicit, ideally avoid them, and responsibly using such apps requires being aware of them.

Finally, it might go without saying nowadays, but AI health apps will certainly raise ethical issues of privacy, especially if data needs to be sent and computed "in the cloud" instead of working only locally on the user's own device and not shared elsewhere.[23] Hence, even if AI health apps' paternalistic influence on users were beneficial for them and did not undermine or diminish their autonomy, such apps might still end up being considered ethically problematic. Consequently, developing possibly paternalistic AI health apps should also include a *privacy by design* approach[24] in order to ensure ethical acceptability in this regard.

However, while the importance of the abovementioned ethical issues can hardly be denied and addressing them convincingly both in normative ethical debate and in the future design of AI health apps is undoubtedly crucial, one can also hardly deny the potential benefits of AI health apps' paternalistic influences on users' health and overall well-being. If designed in an ethically responsible way, (some forms of) AI health apps' paternalistic influence might well be considered ethically acceptable.

4 | CONCLUSION

I started with the assumption that AI health apps may be considered sufficiently autonomous actors when it comes to influencing the users' behavior for their own good. If so, I have contended that they gain a paternalistic potential. Taking the traditional definition of paternalism as a starting point, I have argued that a slightly revised version of it can meaningfully be adapted to the novel phenomenon of AI paternalism in health apps: *AI health apps are capable of showing a paternalistic goal-oriented behavior toward users in that they (a) might interfere with their liberty or influence their autonomy, (b) of which users might not even be aware or at least have not consented to on specific occasions, and (c) do so because of their goal-oriented programming in terms of promoting users' (quantifiable) health*.

Assuming that my conceptual considerations are plausible, AI health apps' paternalistic influence may show as hard, soft, or nudge paternalism. This paternalistic potential leads to a number of ethical issues, notably the (quantified) notion of health employed, the opacity of how particular paternalistic influences are reached and may be justified, the potential of users' infantilization and of undermining their autonomy because of it and because of users' possible "blind" trust in the app, the probability of implicit biases in the (quantified) notion of health, and possible issues of privacy. All of these issues need to be addressed convincingly, both in normative social and ethical debate about health in general and AI health apps in particular as well as when designing such apps.

CONFLICT OF INTEREST

The author declares no conflict of interest.

ORCID

Michael Kühler https://orcid.org/0000-0002-5993-6941

[20]For an overview, see Müller, op. cit. note 3, sec. 2.3.

[21]See Beck, B. (2020). Infantilisation through technology. In B. Beck & M. Kühler (Eds.), *Technology, anthropology, and dimensions of responsibility* (pp. 33–44). Metzler.

[22]See Müller, op. cit. note 3, sec. 2.4.

[23]See ibid: sec. 2.1.

[24]See Schaar, P. (2010). Privacy by design. *Identity in the Information Society*,3(2), 267–274. https://doi.org/10.1007/s12394-010-0055-x; Nordgren, A. (2015). Privacy by design in personal health monitoring. *Health Care Analysis*, *23*(2), 148–164. https://doi.org/10.1007/s10728-013-0262-3

AUTHOR BIOGRAPHY

Michael Kühler is Research and Teaching Fellow at the Academy for Responsible Research, Teaching, and Innovation (ARRTI) at the Karlsruhe Institute of Technology (KIT), Germany, as well as "Privatdozent" (roughly equaling Associate Professor) at Münster University, Germany. His research interests include ethics, metaethics, and applied ethics, especially medical ethics and ethics of technology. Among his recent publications are Fedock, R., Kühler, M., & Rosenhagen, R. (Eds.). (2021). *Love, justice, and autonomy. Philosophical perspectives*. Routledge Studies in Ethics and Moral Theory. Routledge; Beck, B., & Kühler, M. (Eds.). (2020). *Technology, anthropology, and dimensions of responsibility*. Techno: Phil 1. Metzler; and Kühler, M., & Mitrović, V. (Eds.). (2020). *Theories of the self and autonomy in medical ethics*. The International Library of Bioethics 83. Springer.

How to cite this article: Kühler M. Exploring the phenomenon and ethical issues of AI paternalism in health apps. *Bioethics*. 2021;00:1–7. https://doi.org/10.1111/bioe.12886