Philosophy & Technology (2021) 34:1687–1705
https://doi.org/10.1007/s13347-021-00481-4

RESEARCH ARTICLE

Why Digital Assistants Need Your Information to Support Your Autonomy

Jan-Hendrik Heinrichs[1,2]

Received: 30 March 2021 / Accepted: 10 September 2021 / Published online: 23 October 2021

Abstract

This article investigates how human life is conceptualized in the design and use of digital assistants and how this conceptualization feeds back into the life really lived. It suggests that a specific way of conceptualizing human life — namely as a set of tasks to be optimized — is responsible for the much-criticized information hunger of these digital assistants. The data collection of digital assistants raises not just several issues of privacy, but also the potential for improving people's degree of self-determination, because the optimization model of daily activity is genuinely suited to a certain mode of self-determination, namely the explicit and reflective setting, pursuing, and monitoring of goals. Furthermore, optimization systems' need for generation and analysis of data overcomes one of the core weaknesses in human capacities for self-determination, namely problems with objective and quantitative self-assessment. It will be argued that critiques according to which digital assistants threaten to reduce their users' autonomy tend to ignore that the risks to autonomy are derivative to potential gains in autonomy. These critiques are based on an overemphasis of a success conception of autonomy. Counter to this conception, being autonomous does not require a choice environment that exclusively supports a person's "true" preferences, but the opportunity to engage with external influences, supportive as well as adverse. In conclusion, it will be argued that ethical evaluations of digital assistants should consider potential gains as well as potential risks for autonomy caused by the use of digital assistants.

Keywords Autonomy · Self-examination · Self-perfection · Digital assistants

This article is part of the Topical Collection on *Information in Interactions between Humans and Machines*

Guest Editors: Orsolya Friedrich, Sebastian Schleidgen and Andreas Wolkenstein

✉ Jan-Hendrik Heinrichs
j.heinrichs@fz-juelich.de

[1] Institute for Ethics in the Neurosciences (INM-8), Forschungszentrum Jülich, 52425 Jülich, Germany

[2] Faculty of Arts and Humanities, Rheinisch-Westfälische Technische Hochschule Aachen, Aachen, Germany

1 Introduction: Digital Assistants

Ethical analyses of human–machine interaction have predominantly been critical. Ethicists have identified quite a significant number of moral issues which are being raised by the close interaction between humans and machines. These issues have been identified across the whole range of different interactions with machines such as surgical human–machine couplings, discursive interaction between human and artificially intelligent agent, the more classic clicking and typing, and even everyday behavior in which the human participants do not even realize they are interacting with a machine. In the following, I will focus on the middle ground, on the interaction between humans and machines where they are both separate entities, not surgically linked and where the human is aware of interacting with a machine. I am going to focus on the very broad group of systems called digital assistants. This group includes *clippy*, the infamous Microsoft assistant, the recommendation systems on your Amazon website, and Apple's health app, as well as educational software identifying learner types and making suggestions for content and presentation. It is not limited to the systems explicitly named "digital assistant," such as Siri, Alexa, and Cortana. What I do not want to include are social media systems, because these raise completely different issues and provide different, probably less supportive, services.

These systems are ubiquitous. They are not just present in most standard user software, i.e., the apps on your smartphone and the programs on your computer. They also accompany many other human–machine interfaces if the latter allow for user input. Several implanted medical devices come with a remote control or even a smartphone app, which allows for some — often limited — input such as for example insulin pumps providing data on blood glucose and injections and allowing for direct integration with life logging applications. Most devices making up the "internet of things," such as your "smart" fridge, come with such an assistant and so will many future technologies such as self-driving cars or smart-meters for household energy consumption.

Several moral issues are associated with these systems — justice of access (Zuboff, 2015), risk of exploitation of customers and of the people generating data for the training of these systems (Crawford, 2021), algorithmic bias (Christian, 2021), etc. The following will focus on how a common understanding of daily human life changes via its support by digital assistants and how this new conceptualization feeds back into the life really lived. The new conceptualization analyzes a person's daily activity into a number of separate tasks which can be individually structured and optimized, rather than focusing on daily activity as an integrated whole. I want to suggest that this way of reconceptualizing human life is responsible for the much-criticized information hunger of these systems. Understanding this relation might have a direct impact on the moral evaluation of said information hunger.

The article will proceed as follows: after this stage-setting introduction (1), there will be a short overview of the ethical perspective (2) used here including a working definition of the core term autonomy. Thereafter, it will be discussed (3) how digital assistants can aid people in optimizing life's decisions and how optimization processes bear similarities to autonomy as defined here. Because many voices in the debate see digital assistants as more of a threat than an aid to individual autonomy, the following chapter will provide a reply to several such claims (4), whereafter a philosophical reason for suggesting digital assistants threaten autonomy will be identified and refuted (5), before the article concludes (6).

2 The Ethical Perspective: Self-examination and Autonomy

There are a number of ethical dimensions along which the impact of digital assistants can be measured. One dimension would be their impact on individual utility in terms of desire fulfillment.[1] This analysis is far from trivial, in particular because digital assistants do not merely support the satisfaction of desires but also modify and extend the desires which can in principle be fulfilled. This effect — which will play a minor role in the conclusion of this article, too — would have to be measured in such an analysis.

Another important dimension, which in a full analysis should be taken into very careful consideration, is the political one. Digital assistants and the data analysis they enable generate a large amount of power and put it in the hands of actors with little or no democratic oversight or balancing forces on the market (Poon, 2016; Zuboff, 2015).

However, in contemporary literature, there is even more worry about the effects of digital assistants on people's ability for self-determination and autonomy than on their utility (Frischmann & Selinger, 2018; Lanzing, 2016; Maturo & Moretti, 2018; Poon, 2016; Sax, 2021; Susser, 2019, for an analysis of several such criticisms, see "4. Replies to a common objection: reduction of autonomy," below). Complementary to the abovementioned analyses in political philosophy, these critical analyses of digital assistants employ a conception of autonomy focused on self-determination in choice of life's goals and style. I will focus on this line of enquiry in this article.

Before evaluating the influence of digital assistance on autonomy, let me first provide a short working definition for the purpose of this article and differentiate the aspects of autonomy[2] that will play a role in this analysis: self-examination and self-perfection.

First, I take autonomy to comprise an epistemic component and a practical component. The epistemic component can be described with the classical term of self-knowledge. To be autonomous, one must be able to gain knowledge of one's own behavior and to a certain degree one's own mental states including one's desires and emotions, and compare them with one's reflected normative beliefs, i.e., the norms

[1] I will leave out hedonistic theories for the purpose of this article, without thereby implying any judgment on their comparative importance.

[2] The concept of autonomy is one of the most contested in philosophy. I will not go into a detailed discussion of this concept here, but merely provide a working definition. For a good starting point to the scholarly debate about the concept of autonomy see John Christman's *The Politics of Persons* (Christman, 2009) and for a historical account see Jerome Schneewind's *The Invention of autonomy* (Schneewind, 1998).

one takes after reflection (DePaul, 1987) to be valid. In addition, one needs to be able to evaluate one's normative beliefs according to some further standard such as consistency. The practical component is at least partially captured by the notion of self-perfection. To be autonomous, one must be able to adjust one's own behavior and possibly one's emotions to one's normative beliefs.

These two components interlock in a complex way with each other. In contrast to how this description might make it sound, this is neither a straightforward cyclical nor a fully transparent process. Rather, the different epistemic and practical subprocesses often occur in parallel and sometimes involve influences which the person does not — and in some further cases cannot — have conscious access to.

Second, I take autonomy to comprise an active and a passive component. The active component, also called "competence condition" (Christman, 2004, p. 155; cf. Meyers, 2005), consists of obtaining and employing a set of skills and abilities in order to gain self-knowledge and realize self-perfection. If one focuses on this active component, autonomy is something people do.

The passive component, also called authenticity condition (Christman, 2004, p. 155), consists of a certain success brought about by the active component, i.e., by obtaining and employing the abovementioned set of skills and abilities and by the circumstances in which the person lives and acts. This passive component has aptly been described by Christman as obtaining when a person "could come to embrace (or at least not be deeply alienated from)" (Christman, 2009, p. 227) her conditions and the environment that shapes her choices. Focusing on this passive condition, autonomy is something that happens or occurs to people.

That autonomy in general and self-examination and self-perfection in particular are valuable components of individual life is only rarely contested in ethics. This claim resonates particularly strongly with several major strands in contemporary culture, among them are not just the self-tracking and quantified-self movements, but it is quite plausible to understand awareness and attentiveness movements as focused on self-examination. If digital assistants are — as is often claimed — a tool which enables more people to examine their life and live it and shape it according to their own normative beliefs, this seems to be morally welcome.

3 How Digital Assistants Aid in Optimizing Life's Decisions

I want to suggest that three interconnected aspects of the way digital assistants are designed result in some major morally salient effects on the way people lead and self-direct their lives. These effects together make up a major change which requires ethical scrutiny. These three aspects are (1) an optimization model of lifestyle choice, (2) a version of function creep, and (3) increased information dependence of pursuing a lifestyle. The latter issue, the increased information hunger, has often been analyzed as a more or less separate moral issue of machine interaction. In the following, it will be argued that this information hunger needs to be addressed within the larger picture of the reconceptualization of life's decisions as an optimization process.

3.1 The Optimization Model for Digital Assistants

In design thinking for technical assistants, we usually conceptualize life as a set of optimization processes. Depending on the structure of the assistant system in question, it is a single or multi-goal optimization. In most cases, design is for just one goal: physical fitness, job productivity, health, cost-efficient shopping, specific types of pleasure such as musical or literary enjoyment, etc.[3] Only few assistant systems include diverse goals within a single system (Liono et al., 2019). Inclusion of multiple goals usually happens through the user's employing several different tools, such as several different digital assistants for different purposes.[4] Each of these systems optimizes for one (or more) specific purpose(s), and the user somehow integrates these different optimization results into his daily schedule. This decision to conceptualize life as a set of optimization processes has a number of very relevant repercussions on how we are able to live a life assisted by these systems.

First of all, the use of digital assistants forces users to make their decision processes explicit and thereby transparent to themselves. This results from how the structure of optimization systems relates to the decision processes they are created to support. Very roughly, the design of digital assistant requires the definition of a task domain; the option for the user to set — and later on to retain or revise — a goal within this domain; identification of some means, be it by the system or by the users themselves; and tracking of goal pursuit and goal achievement. This aligns fairly well with the simplified structure of autonomy from above. It maps closely to the sub-processes of autonomy, namely monitoring one's own behavior and one's preferences, comparing them with one's normative beliefs, adjusting one's own behavior, and revising or retaining one's own normative beliefs. There are, however, core differences: some processes, which were rather opaque in unassisted everyday decision-making, need to be made transparent within the design of an optimization system, especially goal definition, retention, and revision. And while the components of autonomous behavior occur in parallel, they typically follow a more or less cyclical order within an assistant system. Thus, the use of digital assistants subtly modifies everyday decision-making by requiring us to make some components thereof more transparent, to make our normative beliefs explicit, so to say. This modification affects a growing portion of daily activities.

Task optimization was until recently not a general approach to the activities in our daily lives. Rather, it was limited to a few important tasks which we considered had to be optimized. This changes now that we have digital tools for nearly all the activities of daily living and these tools are based on an optimization model. Digital assistants are not without predecessors of course. People have been using passive, analog tools for optimization

[3] As Sax would point out: They optimize for at least two goals: User engagement and whatever the official goal of the assistant is (Sax, 2021).

[4] There is one obvious issue with which we can put aside quickly: a set of independent optimization processes are not guaranteed, might even be unlikely, to generate an optimal result for multiple goals. The user is left with the task of integrating the results of several independent optimization processes. Whether this is more efficient than trying to optimize one's conduct for multiple goals by hand is at least not obvious.

for quite some time, a palpable example being training journals for diverse types of sport, laboratory notebooks, or similar devices. These predecessors were, however, neither as ubiquitous nor did they allow for a comparatively effortless tracking of goal pursuit and achievement as digital assistants. What had been the exception now turns into the norm of everyday living. The optimization model for daily tasks not only becomes more ubiquitous in the life of particular individuals, but it also enters the lives of people who did not optimize beforehand. The former merely seize new tools for established ways of living; the latter come across these tools via clever marketing activities and by playing around with these tools come to change their way of going about daily activities.

The degree to which this approach to daily activities is new shows up in the problems of a recent data collection project. Among technical problems of collecting task-progress data from participants, Liono et al. encountered a lack of familiarity with the concept of a task itself. They report "Although the instructions during the intake gave an explicit definition of task as»the set of actions/steps/activities needed to reach a particular goal" and provided examples (e.g.»staff meeting for project X«,»designing webpage for client Y«), weekly meetings were extremely important to ensure that participants annotated their tasks in a consistent way." (Liono et al., 2019, p. 5). People seem to only now get slowly used to seeing life's activities as tasks to be tracked and optimized.

Conceptualizing the pursuit of life's goals as an optimization process results in the two abovementioned additional effects of possible function creep and increased information requirement. These effects are not absolutely necessary consequences of the optimization perspective, but they are conditionally necessary given contemporary means of optimization. If our knowledge of optimization processes were different and optimization could be realized with different means, these effects might well not occur. For the foreseeable time, they do, however, seem to be fairly stable.

3.2 Function Creep and Techno-social Engineering

An issue close to the heart of how the optimization model affects our life has been discussed under the term "function creep."[5] According to Koops: "[…] function creep can be defined as *an imperceptibly transformative and therewith contestable change in a data-processing system's proper activity*." (Koops, 2020, p. 1). There seems to be a version of this phenomenon occurring in digital assistants: we start out using an assistant for one narrowly circumscribed purpose, such as step counting to increase our physical activity. But then we figure out, either by ourselves or prompted by the systems design, that it can be used for related purposes too. It can for example also measure how long and how restfully we sleep. Soon we make these purposes our own, sometimes without having given them sufficient previous thought.[6]

[5] The discussion of what exactly function creep is is still ongoing. The term it often defined implicitly in the literature. Koops provides the first detailed analysis of the concept (Koops, 2020).

[6] A similar issue has been raised in 1977 by Winner (1977, p. 238). Under the heading of reverse adaptation, he described how technological systems reverse the adaptation of means to goals as described in theories of instrumental rationality into an adaptation of goals to the means present in the technology. That electronic assistants incite goals in their users by providing the means to these goals is probably one central aim of market-savvy designers.

This is a peculiar case of function creep insofar as it is limited to the use of a technology by a single person across time. Most of the time, “function creep” refers to a process of extension of use or purposes across a community, which a part of the community takes to be contestable. However, if the digital assistants’ proper activity depends on the specific use pattern of individual users, the process described above can be subsumed under function creep. The open question remains, whether the users themselves have a reason to contest the new proper activity of the system, even though they themselves initiated the transformation in question. This depends — among other factors — on the process by which they came to adopt the new function for their device. Core questions in this regard are whether this process is transparent to the individual, whether it is open to revision or refusal, and whether and to which extent there is pressure involved.

3.2.1 Techno-social Engineering as Function Creep

Frischmann and Selinger (2018) are worried about a similar process of function creep, however, not for the individual but the collective case. They develop the concept of “techno-social engineering creep.” As the authors acknowledge, techno-social engineering is an inevitable fact of human civilization. Humans have engineered their social system by technological interventions at least since the advent of agriculture. Many of the decisions about the future of society had and still have the structure of techno-social engineering creep. What Frischmann and Selinger work out in detail is how such creep phenomena often result from situations of choice similar to dilemmas of social rationality. Individually rational decisions, such as using the common pasture, installing advertisement-financed software, or replacing a training diary with a fitness tracker, result in collectively irrational and disputable results such as a loss of the commons or of our right to privacy. They introduce a number of particular versions — surveillance creep, outsourcing creep, boilerplate creep — of this general phenomenon. As Frischmann and Selinger are more concerned with the effects of the use of digital assistants on specific rights and social rules, I will not go into more detail of their fascinating work here. I will, however, make liberal use of their core idea, namely that by designing technical systems we are engaging in techno-social engineering and often do so unaware of the effects. The type of engineering I want to highlight here is not engineering of social systems, but of the way we understand individual life.

3.2.2 Causes of and Reasons for Function Creep

Function creep in individual use mostly does not occur by itself; it needs the opportunity for adopting new functions of the tool in question. Such opportunity is, however, rather common in all but the most specialized tools as it is in most institutions (function creep tends to be a risk for institutions as much as for tools). There are different reasons for designing tools in a way which promotes function creep. The most trivial reason is economic: Designers and programmers work in a corporate environment, which incentivizes data collecting effects of software (Sax, 2021; Zuboff, 2015). They often design software in a way which promotes function creep in order to gain a larger dataset from their users. While the jury is still out for the moral evaluation of this type of design, it seems primarily to raise ethical issues which are familiar from previous contexts (cf. Habermas, 1991; in particular the so-called colonization of the life-world by technical imperatives, already discussed for the case of social media by Van Dijck, 2013) and which for this reason will not be discussed in this article.

The for current purposes more interesting reason for designing software with potential for function creep lies in genuine intent of beneficence: many life goals are closely interwoven. Success in one depends on the effort invested and success realized in the other. In addition, life goals (or interests) often do not come alone, but they build clusters. Someone who learns the piano will likely be interested in specific types of music, and might be interested in home design and furniture (because pianos require some space and are therefore more often found in slightly larger dwellings, which in turn provide opportunity for design and furniture). Someone who learns foreign languages will likely be interested in travel literature and film, etc. Most prominently, health goals are very closely interwoven and optimizing for one often does not just allow for but requires regard for the other. Designing tools which make such interdependencies transparent and offer support in solving not just one optimization problem but in taking account of its relation to others can well be intended as of great utility for users. It is suited to make users aware of interdependencies of the pursuit of their goals, which they might otherwise have overlooked. If that is the case, and there are nevertheless potential losses of autonomy, that would be a dilemmatic moral situation. It would be dilemmatic because the primary effect of getting to know interdependencies of one's preferences and their pursuit is a form of empowerment in self-determined living.

3.3 Optimization and Information Requirement

Due to function creep, the number of optimization tasks in our everyday decision-making grows. And with each optimization task, the amount of data required increases because optimization itself is data hungry. While this happens more in certain types of platform, namely in machine learning systems, it does happen in all types of digital assistants. The data required for the effective use of digital assistants increases for simple and good reasons which have to do with the type of service provided. The more customized to the user (a), the more specific (b) and the more stable across use cases (c), the more information is — ceteris paribus — required for the provision of services.

(a) *User customization* of the machine performance requires more information because only services that do not consider personal characteristics require relatively little information about the user. If the performance of an assistant is to be more relevant to the particular user than such generic performances, then additional information is required. Compare the amount of data required for either the generic services, for example “a book” or “a walker” or “a leg prosthesis” or the user-customized services: “a book in large print or on a font-adaptable reader or as an audiobook," "a walking stick for large left-handed people," "a prosthetic leg for a specific anatomy, physiology, etc."

(b) Information requirement also rises with the desired *specificity of performance*. Machine services can be generic not only in the sense that they are not customized to the user and the conditions of service delivery. They can also be generic in the sense that the exact characteristics of the performance itself are not specified. "A book" is less specific than "a hard science fiction novel with a female protagonist in which uplifting is discussed," "a walker" is less specific than "a shock-absorbing forearm crutch with a closed arm cuff for disabled sports," and "a leg prosthesis" less specific than "a leg prosthesis for alpine mountain climbing".

(c) The higher the desired *environmental stability* of the performance, the more information will be needed, too. The more stable a particular performance is to be, even under widely varying environmental conditions, the more possible disturbing factors must be known in advance in order to be able to compensate for them. The compensation for known unknowns is much simpler — and thus more reliable — than that of unknown unknowns. Thus, the best way to tailor a service to a user is to collect information about the intended conditions of use beforehand. "A book" is in this sense less specific than "a book readable underwater and on a flight," "a walker" less specific than "a walker suited for use on hardwood floors and concrete," and a leg prosthesis is less specific than "a prosthesis useable in cold climate."

Each of these three reasons is based on efforts to optimize a service for a user. The same effect that has here been demonstrated for books, walkers, and prosthesis occurs for digital assistants. The more is known about users, their intended goals, and their environment, the better a service can be customized. The optimization of the service owes its optimality condition to the desires and projects for which the system is being used. Thus, the data hunger of digital assistants is not in the first place a nefarious undertaking of our corporate overlords or of some surveillance state.[7] It is based on generating a service which takes our interest in realizing and expanding self-determination into account.

Why would one think that a digital assistant's gathering information for these reasons might increase our autonomy? I have tried to make this idea salient by describing both digital assistants' optimization model and autonomy in similar terminology. What I hope to have plausibly suggested is that there is a strong analogy in the structure of deliberate optimization tasks and in autonomous decision-making. Furthermore, digital assistants are suited to compensate for human weaknesses in autonomous decision-making. First, they urge users to specify goals fairly precise

[7] This is not to deny that there are political and economic reasons for extreme data collection processes. These have been discussed in detail by Zuboff (2015). The reasons discussed above and those analyzed by Zuboff are not mutually exclusive. The data hunger of digital assistants seems to be causally overdetermined.

and explicitly and use the data collected by the devices to track goal achievement.[8] By making these goals explicit within the digital system, they allow for more reflective goal revision or retention. By making use of the information for goal achievement tracking, they support another activity which humans often struggle with, namely objective, information-based self-assessment.[9]

The above is intended to demonstrate that there are justified reasons for the data hunger of digital assistance. Critiques off the data collection by digital assistants in general or a particular system should therefore take care to differentiate between information gathering for these and similarly justified reasons and information collection going above and beyond such reasons. While the latter surely are a deserving target of concern, the former should be handled with more care.

4 Replies to a Common Objection: Reduction of Autonomy

Given this reason for the data hunger of digital assistants, here is a disconcerting observation: in the ethical literature, it has predominantly been claimed that rather than supporting autonomy digital assistants are suited to reduce the autonomy of their users.

The — admittedly oversimplified — argumentative structure of several articles critical of the autonomy supporting and empowering potential of digital assistants is the following: digital technologies, be it assistants or choice architectures or AI systems, promise means for supporting our self-determination. However, because these tools themselves or the decisions we take with the support of these technologies have a specific shortcoming (they are for example not audience-specific like a diary, they are invisible to us, they involve a pathway for manipulation (see below)), they are not really an improvement of self-determination. Quite the opposite, they are means of endangering our autonomy. Here are a number of quite convincing warnings about the dangers to autonomy posed by digital assistants which employ one or the other version of this argumentative structure:

Lanzing (2016) compares self-tracking devices, especially health trackers, to older, analog techniques of self-tracking such as diaries. She diagnoses that digital self-tracking devices, unlike a diary, tailor and possibly expose information about the user to an unspecified audience. A diary or a written report about one's physical activity tends to have a specific audience, and the information therein can be tailored to that audience. Digital tracking devices on the other hand always raise the possibility of the collected information being read by a number of different audiences, some of which the user would not want to gain access to that information, many of which

[8] Admittedly, there is a huge influence of individual psychology on how people react to their goals having been made explicit. For some it might have habituating effects, making it unnecessarily difficult to revise their goals.

[9] As an anonymous reviewer pointed out, the categories available for self-assessment by a particular digital assistant have — in most cases — not been generated by an objective and neutral process. Thus, while the assessment might be objective, it need not be — and probably rarely is — neutral or unbiased.

have an agenda of their own. For this reason, self-tracking devices are, according to Lanzing, not reliable means of empowerment, but risk autonomy.

Susser (2019) diagnoses not just digital assistants but all kinds of digital choice architectures as often invisible to the user. People become so used to using a specific technology for their purposes that they do not see the technological interface anymore, just the task they are putting it to use for. The technology becomes — to use the Heideggerian term — ready to hand. It turns invisible and by turning invisible it allows for manipulation. We lose the ability to see how it influences our behavior and therefore cannot critically reflect on it anymore. Only if we come to perceive the technology itself again are we able to see that it influences our choices. Susser refers to the nudge framework of Thaler and Sunstein (2008) for his diagnosis, and he insists that this kind of influence can only be autonomy preserving if it is made transparent.

Sax (2021) draws our attention to the fact that in particular health apps optimize not just for the health of individuals, but for user engagement. They are trying to look to convert users into paying customers, and thereby are prone to manipulating the users towards ends which they do not necessarily share — namely spending money on the app. The means employed for that goal, and here Sax refers to Susser, remain hidden from view in most, if not all, cases. Like Susser, Sax identifies the threat of manipulation of a user's decisions as the main threat to autonomy.

Many of the warnings regarding the shortcomings of decisions made with the support of digital assistants are quite adequate. These technologies do involve a threat of manipulation; they are intended to grab the attention of users and to gather more information about them, which can be marketed (for a detailed and fairly balanced analysis of these issues cf. Maturo & Moretti, 2018). I do not want to deny that this is a process which is likely to occur in the use of digital assistants.[10] What I want to argue however is that (4.1) the risk of a loss of autonomy is derivative to a potential for gain in autonomy, and (4.2) postulating such a loss of autonomy seems to credit the digital assistance with more agency than they actually have.

4.1 Loss of Autonomy to Digital Assistants Presupposes Possible Gain

To show how the risk to autonomy is derivative to potential gains, it helps to distinguish cases of people with low and with high previous levels of autonomy. Persons are limited in their autonomy or self-determination if they lack the ability to set their own goals, to promote them, or to monitor goal achievement — be it self-imposed or due to their circumstances. There can be several different causes for a low level of autonomy, such as insufficient skills, abilities, or resources for self-determination or

[10] There are two non-exclusive options to deal with this diagnosis: option number 1 is to identify possible means of mitigating the risk to autonomy. This will have to be done on an institutional level. The main options probably are regulations for advertisements/marketing as already established in many countries. Option number 2 is to balance the potential benefits against the risks and pick digital assistance only if there is a net potential gain. This rests on the individual level alone and — paradoxically — is a case of autonomous choice.

for managing external influences on their decision processes. While there might be exceptions, these persons are unlikely to lose autonomy to a digital assistant because the main ingredients of autonomy are not at their disposal to start with. If persons set few of their purposes in life themselves and instead adhere to the goals and means set by others, they will not lose autonomy by using digital assistants. Rather, they will move the locus and means of control from other people or systems to digital assistants. If persons are not able to track their goal achievement themselves, they will not lose this ability by relying on a digital assistant. Imagine persons who have never engaged in any kind of physical exercise beyond what they have been forced to, say in school, who have not taken control of their bodily fitness, not even in the form of intentional neglect. It can hardly be said that such persons lose existing self-determination in this regard if they turn towards some fitness app. Such persons do not lose self-determination if they unquestioningly follow the digital assistant's suggestion for workouts and nutrition. So, if one were to claim that the digital assistant is a threat to these persons' self-determination, one would have to show what exactly they might lose. I want to claim that there is only one possible candidate for this: what is lost is not existing self-determination — because there is little or none — but a part of the potential for self-determination *gained* by turning towards the systems. These persons could engage with the assistant in a more competent, more self-determined way. They gain the potential, the possibility to determine herself in questions of bodily fitness. However, according to the argument they immediately lose this potential again because of the allegedly autonomy-endangering nature of the digital assistant.

In a nutshell, for the risk of losing autonomy to digital assistants to be plausible in this population, there first has to be potential for gaining autonomy which is then to be lost.[11] The question is whether the gain in potential autonomy might outbalance the risk of losing this potential. Even if the probability that this gain is sustained is very low, there still is a net gain in potential autonomy left.

There seems to be more of a threat to the self-determination of people who already have a high degree thereof. People who regularly govern themselves, who set their own goals, determine the means to these goals themselves, and monitor whether they achieved their goals seem at risk of losing this degree of autonomy by turning to digital assistants according to the arguments introduced above. However, this seems counterintuitive. The people who are most used to setting their own goals, to monitoring them, and to picking means suited to pursue those goals are thought to fall into a fairly obvious trap and lose exactly these abilities. This implies a high opinion of the lure of digital assistants. There are several possible forms this lure can take, such as simple convenience among others. But one of the more tempting lures for the persons in question seems to be the promise of a gain in autonomy.

[11] Please note that there are digital systems which simply replace the heteronomous direction of people's lives. Such systems do not just give persons exact directions, but also track whether and how fast they follow those directions; they are connected to some sort of sanctioning mechanism, be it report to the management or direct link to a payout function. It would, however, be strange to call such systems digital assistants.

Digital assistants at least seem to these individuals to enable them to either set their goals with more deliberation, to monitor their achievement with more precision or as means of pursuing goals more efficiently. Thus again, the risk of losing autonomy to digital assistants seems to be derivative at least of the promise of gaining autonomy.[12]

Admittedly, most critical authors accept that digital assistants have a potential to be empowering, to improve the autonomy of the agent. What they rarely make explicit is that the risk of losing autonomy depends on this potential. Given that these devices nearly always combine potential benefits and risks for self-determination, a thorough ethical analysis would need to weigh these against each other instead of basing ethical advice on the risks alone.

4.2 The Model for the Loss of Autonomy to Digital Assistants

The very idea that we are losing autonomy to a machine seems to be derivative to another case of losing autonomy, namely of losing autonomy to other people. Losing autonomy means losing the ability to govern oneself, and the only option for this is to either become governed by someone else or not to be governed at all.[13] The former implies that there is some agent who determines my behavior; the latter implies that my behavior is not governed but determined by something else without this determining force having any normative power.

The idea that we might lose autonomy to digital assistants can be associated with both options. According to the first option, the digital assistant either has a degree of agency sufficient for governing someone's actions or it is someone else's tool for governing user's actions. According to the second option, the influence on the user's behavior merely looks like some kind of governance. But the digital assistant lacks any obligating force and following its prompts is a case of being determined by mechanisms without any normative content. The human behavior prompted by the assistant in turn becomes more similar to that of a simple machine. This idea has been followed up in detail by Frischmann and Selinger (2018). While this is an interesting line of enquiry, I will ignore it for the purpose of this article and return to the idea that the digital assistant is either itself a governing agent or the tool of such an agent.

The first variation of this objection is that the digital assistant itself has sufficient agency to govern the user's behavior. This, however, seems to misinterpret the nature of digital assistants. There currently is no evidence that digital assistants can be agents in a sense remotely similar to the way humans are agents. One of the least metaphysical exuberant reasons to think otherwise is that these devices are best explained from an intentional stance (Dennett, 1971). We do best in our

[12] If the latter scenario were plausible, there would still be some balancing required: as discussed there is some potential gain in autonomy for those who currently have little thereof, while there is a potential loss in self-determination in the more autonomous persons.

[13] In addition, in the original Kantian version in the *Critique of Practical Reason*, one loses one's autonomy to one's passions if one comes to be determined by them and not by reason. Given Kant's account of the passions, this is fairly close to the latter version above, namely losing autonomy to an influence without normative power.

prediction of digital assistants if we ascribe mental states, beliefs, and intentions to them. While there are good reasons to interpret these devices and their output as intentional systems and intentional states, taking an intentional stance is an explanatory strategy, not an imposition with normative relevance. One can at the same time consider a device an intentional system and a mere tool.

The other option, namely that digital assistants are tools for governing people, would require determining who the governing body is. As far as I can see all the options for an answer to this are either contentious (the programmer, the tech-corporation) or lead back to some kind of self-determination, if not necessarily the traditional atomistic version (society as a whole, a collective agent comprising user, programmer, distributor, etc.).

This is a major strand in contemporary debate about the role of the large technology corporations in political and market structures, which would require far more attention than I can give it in this article. Let me therefore only briefly say why I think that the claim we might lose autonomy to the programmer or company behind a digital assistant is contentious. The idea mentioned in a number of books and articles (in particular Zuboff, 2015) is that by manipulating or rather by designing the choice architecture of an application, the designers or the corporation behind them can manipulate a person into doing something, most of the time into spending more money on their products. However, people arranging a choice architecture in a way that serves their own ends are in most cases not infringing on our autonomy; they might make it harder to stick to certain austere preferences, while they support and, in some cases, generate other more indulgent preferences. The anathema of autonomy is domination, not temptation.[14] Anderson (2014, p. 137) explicitly claims that "[a]utonomy competencies include, for example, the ability […] to actually carry out one's intentions in the face of temptations." While I do not want to deny that some such corporations gain enormous market power, maybe more than modern democracies should grant them, this alone does not suffice to turn their use of digital assistants, and their data gathering and analysis into a danger to our autonomy.

5 The Comparison Class: Ideal Choice Architectures and True Preferences

Why, however, are digital assistants so often seen as a threat to our autonomy? In the following I want to explain, why — I think — many authors unilaterally stress the autonomy-undermining character of digital assistants and seem to overlook or underrate their autonomy-supporting potential. I want to provide an error theory, so to speak, and argue that there is a philosophical reason for this temptation. This temptation closely relates to the second differentiation I made to autonomy above, that between an active and a passive component. Digital assistants start to look

[14] This is not to exclude that some extreme forms of temptation, for example those which generate unforeseeable addictions, which in turn can only be fed by a small circle of agents, might come close to domination.

threatening if one over-emphasizes the passive component of autonomy. Let me make this distinction clearer with an instrument from classic analytic philosophy, the distinction between action and success terms (Ryle, 2013, p. 149).

5.1 Autonomy: Action and Success

"Being autonomous" can be read either as a description of an action or process — as an action term — or as a description of a resulting state, a success or achievement term so to speak. As an action term, it refers to the abovementioned active component of employing a set of abilities and capacities to determine one's own life in direct engagement with one's environment. As a success term, it refers to the passive component, a property of one's completed decisions, namely the property of not having been influenced by factors which the person would not endorse (or be deeply alienated from) if they were transparent to her (see above (2) and Christman, 2009, p. 227).

To be more precise, "being autonomous" is neither just an action nor just a success term; it comprises an active and a passive component. You cannot really be autonomous, just by having the success of not being influenced by anything that you would not endorse. If you are in the lucky position of not being influenced by anything of that kind, and did not employ any of the capacities, skills, and abilities that typically make up the autonomous leading of your own life, it is hard to consider you autonomous. You are just lucky. If on the other hand, you employ all these skills and capacities, and still fall prey to a form of manipulation you would not endorse, we would not call you autonomous either. Neither an action nor a success term alone seems to be adequate. Being autonomous seems to be both, a type of action and a type of success.

How are these considerations about the term autonomy related to explaining why authors unilaterally stress the potential losses in autonomy caused by the use of digital assistants? I want to claim that by putting too much emphasis on the success component of autonomy, one can come to ignore that being autonomous is something people do. We come to think that autonomy is a success which can be brought about by ideal circumstances such as ideal choice architectures. This picture removes the agency of the autonomous person from view. If autonomy is a success to be brought about by the circumstances, then this seems to raise impossible standards for digital assistants' and all other choice architectures' effect on people's autonomy. If, in turn, digital assistants are measured against standards, which only ideal choice architecture could fulfill, they will unilaterally be seen as threats to autonomy. The positive effects which might show when measured against different, more realistic standards drop from view.

5.2 On the Tacit Assumptions on the Comparison Class

Reading through the literature warning against a loss of autonomy to digital assistants, it springs to attention what current design practices and their results are being compared to. Indeed, it seems as if current design has to compete against ideal choice architectures preserving or promoting true interests.

Let me substantiate the claim with a short glimpse at the literature. Sax (2021, p. 14) for example insists on a duty to care for for-profit health apps and suggests that they infringe against this duty if they "*systematically and indiscriminately* optimize for engagement, retention, and conversion," that is, if they do not exempt people with health vulnerabilities from their marketing proposals. Sax's duty to care is intended to counter the threat of manipulation and "[c]ases of manipulation are characterized, then, by the manipulator's disregard for, or indifference to, the manipulee's true interests." (Sax, 2021, p. 8).

A similarly strong requirement is formulated by Susser who suggests that "[t]hose filling our everyday decision-making environments with adaptive choice architectures must be entirely forthcoming, not only about the fact that they are using such tools, but about their purpose (intended outcomes), and about the mechanisms by which they achieve it." (Susser, 2019, p. 5 f.). This kind of requirement is typically not imposed on other agents who generate choice architectures for our everyday decision-making. Again, the standard for the choice architecture is the support of the user's interests: "Creators of adaptive systems ought to be held to the highest standard of care—they should be required, that is, to create systems that beyond any doubt advance users' interests." (Susser, 2019, p. 6). Susser is less forthright with the formulation, but what he refers to seem not to be the interests as they emerge from the interaction between the user and the choice architecture, but a set of original, true interests.

Again, I do not want to deny that extra precautions for highly influential actors in society are politically prudent. However, the idea behind some of these claims seems to be that the choice architecture in question should support the users' individual autonomy by not distorting their preferences. This in turn seems to presuppose that for each user there is a set of undistorted preferences, the unhindered pursuit of which constitutes this person's autonomous life.

If digital assistants are measured against a standard according to which choice architectures must be designed in order to promote the undistorted, true preferences of decision-makers, they cannot succeed. They cannot succeed because this standard is not merely unrealistic; it is based on an erroneous conception of preferences.

Preferences are not a fixed quantity that exists prior to the establishment of choice architectures. Rather, which preferences a person has and which he or she pursues always depend on such architectures (Thaler & Sunstein, 2008, p. 75 f.). In this respect, there are no neutral or ideal choice architectures to which one can simply compare digital assistants. Which preferences can be considered autonomously set cannot be evaluated before but only after a person interacts with a choice architecture. It does not depend on whether any supposedly true preferences or unmanipulated normative beliefs have been preserved, but on whether the person has been able to use and has used the skills that belong to the active component of autonomy, and whether he or she feels alienated from the outcome.

Thus, overstressing the autonomy-endangering character of digital assistants seems to have a deeper philosophical reason: it is caused by formulating a standard which requires a choice architecture to protect a person's ideal or initial preferences. And such standards in turn seem to me to be based on too much emphasis on the success component of autonomy. Following this emphasis, one might think autonomy can be realized by optimizing persons' environment for transparency and non-interference with their decision-making. And digital assistants do not do this; they alone do not realize their user's autonomy. No real choice architecture does. As Anderson puts it in an enlightening simile: "In this sense, being autonomous is like being able to find one's way through the woods: you have to discern where you want to go, figure out how to get there, persevere through the brambles, and occasionally stop to ask yourself whether the trip is worth the effort." (Anderson, 2014, p. 137). If there are no brambles, one does not have the opportunity to be autonomous.

6 Conclusion: Loss of Autonomy and the Breakdown of Tools

What I hope to have shown is that digital assistants can make a profound change to the way people understand their life. By making a specific type of tools available for everyday activities, the mode of engaging with these activities is being changed. People come to see daily activities as something to be optimized. This change in perspective and tools is the dominant reason for the data hunger of digital assistants. The optimization model and the resulting tools raise the potential for improving people's degree of self-determination. They do so because the optimization model of daily activity is genuinely suited to a certain mode of self-determination, namely the explicit and reflective setting, pursuing, and monitoring of goals. Furthermore, the newly intensified generation and analysis of information supports self-determination by allowing for objective and quantitative self-assessment.

In trying to answer some common objections, I hope to have shown that the fear of digital assistants reducing their users' autonomy tends to ignore that the identified risks to autonomy are derivative to potential gains in autonomy. Thus, ethical advice should at least be based on a comparison of potential benefits and risks to autonomy and not on the latter alone. Furthermore, I tried to explain why many authors raise such extensive worries about the autonomy-reducing effects of digital assistants. I found one reason in an overemphasis on a certain aspect of autonomy, namely autonomy as a success independent from the effort of the agent. I suggested against this that being autonomous does require engaging with external influences, supportive as well as adverse.

The conclusion from the above diagnosis surely cannot be to trust our digital devices blindly, nor to refrain from social and legal precautions against intransparent or even deceptive software. Rather, I want so to suggest that we do not lose sight of the potential for supporting self-determination, for transparent self-examination and effective self-perfection such devices harbor, *if* they adhere to certain standards of transparency. Rather than basing our ethical evaluation and subsequently our social policies on reactions to risks alone, we should take into account the potential benefits, as well. And the potential benefits of these devices seem to lie in enhancing one of the most important human traits, the power for self-determination.

Data Availability NA

Code Availability NA

Declarations

Conflict of Interest The author declares no competing interests.

References

Anderson, J. (2014). Autonomy and vulnerability entwined. In C. Mackenzie, W. Rogers, & S. Dodds (Eds.), *Vulnerability. New essays in ethics and feminist philosophy* (pp. 134–161). Oxford University Press.

Christian, B. (2021). *The alignment problem: How can machines learn human values?* Atlantic Books.

Christman, J. (2004). Relational autonomy, liberal individualism, and the social constitution of selves. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition, 117*(1/2), 143–164

Christman, J. (2009). *The politics of persons: Individual autonomy and socio-historical selves*. Cambridge University Press.

Crawford, K. (2021). *The atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press.

Dennett, D. C. (1971). Intentional systems. *Journal of Philosophy, 68*(February), 87–106.

DePaul, M. R. (1987). Two conceptions of coherence methods in ethics. *Mind, 96*(384), 463–481.

Frischmann, B., & Selinger, E. (2018). *Re-engineering humanity*. Cambridge University Press.

Habermas, J. (1991). *The structural transformation of the public sphere: An inquiry into a category of bourgeois society*. MIT Press.

Koops, B.-J. (2020). The concept of function creep. *Management of Innovation eJournal.*

Lanzing, M. (2016). The transparent self. *Ethics and Information Technology, 18*(1), 9–16. https://doi.org/10.1007/s10676-016-9396-y

Liono, J., Trippas, J. R., Spina, D., Ra-haman, M. S., Ren, Y., Salim, F. D., . . . White, R. (2019). *Building a benchmark for task progress in digital assistants.* Paper presented at the Task Intelligence Workshop at WSDM 2019, New York, NY.

Maturo, A., & Moretti, V. (2018). *Digital health and the gamification of life: How apps can promote a positive medicalization*. Emerald Publishing Limited.

Meyers, D. T. (2005). Decentralizing autonomy — Five faces of selfhood. In J. Anderson & J. Christman (Eds.), *Autonomy and the challenges to liberalism* (pp. 27–55). Cambridge University Press.

Poon, M. (2016). Corporate capitalism and the growing power of big data: Review essay. *Science, Technology, & Human Values, 41*(6), 1088–1108. https://doi.org/10.1177/0162243916650491

Ryle, G. (2013). *The concept of mind*. Barnes & Noble.

Sax, M. (2021). Optimization of what? For-profit health apps as manipulative digital environments. *Ethics and Information Technology*. https://doi.org/10.1007/s10676-020-09576-6.

Schneewind, J. B. (1998). *The invention of autonomy - a history of modern moral philosophy Cambridge*. Cambridge University Press.

Susser, D. (2019). *Invisible influence: Artificial intelligence and the ethics of adaptive choice architectures*. Paper presented at the Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, Honolulu, HI, USA. https://doi.org/10.1145/3306618.3314286.

Thaler, R. H., & Sunstein, C. R. (2008). *Nudge. Improving decisions about health, wealth, and happiness*. Yale University Press.

Van Dijck, J. (2013). *The culture of connectivity: A critical history of social media*. Oxford University Press.

Winner, L. (1977). *Autonomous technology : Technics-out-of-control as a theme in political thought*. MIT Press.

Zuboff, S. (2015). Big other: Surveillance capitalism and the prospects of an information civilization. *Journal of Information Technology, 30*(1), 75–89. https://doi.org/10.1057/jit.2015.5

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.