Is Meaningful Human Control Over Personalised AI Assistants Possible? Ethical Design Requirements for The New Generation of Artificially Intelligent Agents

Kuilman, S.K.; Nyholm, Sven; Buijsman, S.N.R.; Cavalcante Siebert, L.

DOI
10.1007/s13347-025-00976-4

Publication date
2025

Document Version
Final published version

Published in
Philosophy & Technology

Citation (APA)
Kuilman, S. K., Nyholm, S., Buijsman, S. N. R., & Cavalcante Siebert, L. (2025). Is Meaningful Human Control Over Personalised AI Assistants Possible? Ethical Design Requirements for The New Generation of Artificially Intelligent Agents. *Philosophy & Technology*, *38*(4), Article 148. https://doi.org/10.1007/s13347-025-00976-4

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Philosophy & Technology (2025) 38:148
https://doi.org/10.1007/s13347-025-00976-4

RESEARCH ARTICLE

Is Meaningful Human Control Over Personalised AI Assistants Possible? Ethical Design Requirements for The New Generation of Artificially Intelligent Agents

Sietze Kai Kuilman[1] · Sven Nyholm[2] · Stefan Buijsman[1] · Luciano Cavalcante Siebert[1]

Received: 20 March 2025 / Accepted: 12 September 2025

Abstract
Recently, several large tech companies have pushed the notion of AI assistants into the public debate. These envisioned agents are intended to far outshine current systems, as they are intended to be able to manage our affairs as if they are personal assistants. In turn, this ought to give users a leg up, as one prominent tech exec has put it. However, it remains to be seen how these Personal AI Assistants (PAIAs) are implemented, and critical reflection on how and whether they can be implemented in a responsible way is needed. Currently, such agents are undertheorized and this may cause us to misunderstand their value and capacity. In this paper, we explore and critique the potential for responsible implementation by considering some design requirements based on the notion of meaningful human control. If we desire to have control over such assistants, then we need to be able to do so meaningfully and effectively. In looking at the design requirements, we run into the issue that their broad and differing capacities make any kind of design requirements hard because there are simply no standards to which we can measure PAIAs. Furthermore, it seems that the implementation of these assistants will be a matter of trade-offs both in capacities and in values, which will likely lead to enhancement for some rather than an improvement for all.

Keywords Meaningful human control · Personal artificial intelligent assistant · Enhancement · PAIA · MHC

Extended author information available on the last page of the article

Published online: 18 October 2025

1 Introduction

Google, Microsoft, Meta, and OpenAI have all recently shot their developing ideas of Personal AI Assistants—sometimes simply referred to as "agents"[1]—into the public debate[2]. Supposedly, these agents will far surpass the Siris of yesteryear. Microsoft's Copilot, OpenAI Assistants API, Meta AI, Google's Gemini, and Inflection's Pi, are some of the examples that sprung up in the second half of 2023 (Wiesinger et al., 2024). According to an extensive report on the ethics of these "advanced AI assistants" by Google DeepMind (Gabriel et al., 2024), there is reason to believe that the scope of these systems will broaden, as well as their capacity for action and autonomy. While these new AI agents and personal AI assistants are still in development and the above-mentioned tech companies are vague in their descriptions of what they are aiming for, it is clear that this is intended to be the next big thing within the world of AI.

One main stated aim behind such AI assistants is that they will aid people in ordering their lives and managing certain parts of their affairs. They could also be used by organisations to increase effectiveness. As Gabriel et al. (2024) explain, these agents are envisioned as being able to "plan and execute sequences of actions on the user's behalf across one or more domains and in line with the user's expectations." (cf. Gabriel et al., 2024, section 2). During an interview, Sundar Pichai, CEO of Google, said about this kind of technology that "[i]t may give users a leg up" (Chamberlain, 2024). The proverbial leg up here seems to relate to what philosophers have in mind by the notion of human enhancement (Bostrom & Roache, 2008). Yet, the discussion surrounding AI assistants can also be compared to other (AI) technology in the sense that such technology is disruptive and may alter our ways of thinking or our perception of society and morality (Danaher & Sætra, 2023). The way in which Personal AI Assistants (PAIAs) are expected to be incorporated into our lives raises a range of additional questions. For example, these systems may be taking over tasks involving important moral and legal responsibilities (Milano & Nyholm, 2024; Santoni de Sio & Mecacci, 2021). If these PAIAs are managing our affairs and entering into contracts for us, what will happen if things go wrong? Consequently, part of the debate ought to be about what philosophers call potential blame and praise gaps (Nyholm, 2023). Who do we blame when such AI systems make a mistake? And who deserves credit, when such systems benefit users?

In this paper, we are concerned with control over PAIAs. Control over such systems is highly relevant (even compared to other AI systems), given that we are handing over our personal details and personal affairs to such systems and are supposed to let them act on our behalf in relation to (more or less important) events within our lives, both on a professional and a personal level. Meaningful human control (MHC)—a notion that has been discussed in contexts such as automated weapons

[1] For the sake of this paper, we are concerned with agents that resemble large language models but are likely aided by different kinds of modules for reasoning, as well as having access to different environments (e.g. financial and planning) and information. Although the term agent can mean a variety of things, the exact technical specification of what an agent is remains outside the scope of this discussion.

[2] http://www.ibm.com/think/topics/ai-agents

systems, autonomous vehicles, and more (Mecacci et al., 2024)—may be regarded as a foundational requirement if we want PAIAs to be effective, trustworthy, and reliable as personal assistants. The potential for error, and possible resulting harms, is large and will be a pressing issue for the individuals using such a technology. Based on the concept of MHC, we formulate ethically informed design requirements. We apply and concretize the concept of MHC to the domain of PAIAs and show what problems we can gain insight into and what problems result from the current undertheorization of the general idea of PAIAs. These design requirements can provide a jumping off point to help shape the debate surrounding the problems that may ensue from the development and deployments surrounding PAIAs. What should we want from these technologies (and should we want such technologies at all (Mitchell et al., 2025))? If we desire them to be beneficial, we need to understand what trade-offs may arise from their use and whether they can actually provide the projected benefits tech companies ascribe to them.

Notably, theoretical discussions surrounding the ethics of AI assistants and AI advisors are not particularly new. Gabriel et al. (2024) suggest that such agents can help with planning, gathering information, generating ideas, finding longer-term goals, and even interacting with other humans or assistants. All these concepts have already been discussed separately in the existing (ethics of) AI literature. Similarly, before PAIAs there was already talk of AI coaches (Lara & Deckers, 2020) and advisors (Pisoni & Díaz-Rodríguez., 2023), both of which can provide personalized help and bespoke services to people, just as skilled human assistants could. The problem of figuring out what users actually want (aligning to the right goals) has been discussed by Tielman et al. (2018) in their discussion surrounding decision support systems. There have also been discussions of AI assistants as potential moral advisors (O'Neill et al., 2022) which could aid us in decision-making and cognition surrounding difficult choices. If PAIAs will be able to enter into contracts for us, we can also take clues from automated negotiators and the related challenges, as already discussed in the literature, for instance, in terms of how to update behavioral strategies over time (as our preferences may also change over time) (Baarslag et al., 2017).

The potential uses of PAIAs show that we may need to deal with certain problems also established in automated negotiators, AI advisors, and coaches. Yet, we must also consider the worries that such systems may engender. This is not a completely new discussion. Several years ago, in a previous issue of this journal, Danaher (2018) already proposed a general ethical framework for the ethics of AI assistants, in which he suggests that such assistants may cause cognitive degeneration or disrespect autonomy and skew our interpersonal relationships. More recently, we have seen investigations into how interaction with AI assistants might shape our moral development (Yeung et al., 2023). Some research has also been done on the mediative standing of AI voice assistants (Kudina, 2021) and how it may shape our values. In combining so many activities from previously separate systems (e.g. advisors, coaches, negotiators) and extending their capacities, there is a chance that new problems not seen in these different systems may arise as well.

With the proposed possibilities and extensive use in the personal domain, we need to reflect on what kind of effects these systems can have on the user and whether it is feasible or even desirable to hand over certain types of tasks to AI assistants. As

Milano and Nyholm (2024) already suggest, PAIAs may not be ethically or legally feasible. These PAIA systems are supposed to act on our behalf, but this may cause problems. What if the system makes a choice that affects a user who is not privy to the same information that the system has based its decision on? If an accident occurs at that point, who is responsible? And who is responsible for making sure that such a state of affairs is prevented in the first place? These systems are broadly defined, as noted above, as agents with a natural language interface, which plan and execute sequences of actions on the user's behalf but also in line with the user's expectations. But how do we know what a user's expectations are, and how personalized ought assessment to be? What types of decisions are acceptable (only low-stakes decisions or some high-stakes decisions as well?) to outsource to such a system, and which are not? As we see things, what these questions all have in common is the basic concern to enable meaningful human control over PAIAs.

Our discussion of control and PAIAs is divided into three parts. First, we briefly introduce meaningful human control (MHC). Second, we apply MHC to PAIAs and delve into three different types of issues, duties, personalisation, and idealisation, as well as the relationship between these types of issues. Issues with duties entail the problem that follows from the levels of stakes involved in the use of PAIAs. The problems surrounding personalisation has to do with how much it should be focussed on your gain versus that of others. When it comes to idealisation, we see that there is a trade-off between idealised preferences and exact preferences. In the first case it may be less comprehensible but in the latter we may not aptly capture preference drift or get stuck in suboptimal routines. After delving into the potential problems, we conclude by going over the idea of enhancing the user and investigating how this relates to the differences among users. More specifically, we analyse user enhancement from two perspectives: ubiquitous enhancement, which addresses the question of whether the improvements brought by PAIAs can be considered accessible and beneficial across different groups and actors; and directional enhancement, which focuses on whether enhancement aligns (or not) with what one truly values, taking them in their desired direction.

2 Meaningful Human Control

The notion of meaningful human control originated in the autonomous weapons debate (Roff & Moyes, 2016), with the idea that we have to ensure that humans have control of some sufficiently significant sort over advanced AI technologies intended for use in armed conflicts. The stakes in these cases are incredibly high, life-and-death decisions need to be made, and ensuring human accountability is therefore imperative. Such technology, one can argue, has to be predictable, transparent, and reliable. On the side of the human, this requires accurate information and the capacity to intervene. From those discussions spawned ideas about MHC in other domains, such as automated driving. One influential account is that of Santoni de Sio & Van Hoven (2018). They proposed looking at MHC through the notion of guidance control found in general philosophical debates about free will and moral responsibility (Fischer & Ravizza, 1998). Guidance control itself is based on the premise that people can

be held responsible for actions and the outcomes they produce only if the people in question were guiding the actions in some significant sense, and this is thought to require a moderately reason-responsive mechanism and an ownership condition. What Fischer & Ravizza (1998) mean by this can be summarized as follows. First, a moderately reason-responsive mechanism is a process which instantiates an action based on reasons. Thus, if different reasons apply, the action changes (the "tracking" condition). Second, the ownership condition means that the agent has taken or is able to take responsibility for the action (the "tracing" condition).

The notion of MHC over advanced technologies articulated by Santoni de Sio & Van Hoven (2018) applies versions of these tracking and tracing conditions to the behavior of these technologies and their relation to the humans involved (and also changes the formulation of the conditions somewhat). As Santoni de Sio & Van Hoven (2018) see things, having meaningful human control over a given system requires, first, that the technologies operate in ways that track the relevant moral reasons of the relevant people involved and, second, that it is possible to trace the behaviors of the technologies to individual(s) in the chain of events who understand the technologies and their social significance, and who can therefore reasonably be held responsible for the machine behaviours in question or their outcomes. Note that this does not mean the people in question are directly responsible at that specific juncture in time when an accident happened. A classic example of this is driving drunk. A drunk driver may not be in control over his actions, but he certainly put himself on this trajectory by drinking and getting into the car drunk. An individual can be held responsible for the trajectory they chose. Such a person may be in possession of certain agential capacities with which they could have acted differently so that they are in a general sense in control over how they act and behave in the world.

This tracking and tracing conception of MHC has recently been further applied to other areas than autonomous weapons systems and autonomous driving (Umbrello, 2020 and can be seen as a key component of a responsible way to think about the design of values into AI systems at large (Siebert et al., 2023). Siebert et al. (2023) suggest four defining properties based on the concepts of tracking and tracing to further delineate what the practice of MHC ought to be on the ground floor.

First, the human-AI system has an explicit moral operational design domain (MODD) and the AI agent adheres to the boundaries of this domain. Second, human and AI agents have appropriate and mutually compatible representations of the human-AI system and its context. Third, the relevant humans and AI agents have ability and authority to control the system so that humans can act upon their responsibility. Fourth, actions of the AI agent are explicitly linked to human agents, who are aware of their responsibility.

These four properties are relatively simple to understand. The MODD criterion that is part of the first specified property implies that there is a sociotechnical element to keeping track of the relevant moral reasons. As a condition, AI systems ought to be associated with certain explicit positive and negative duties that apply within the domain in question. Even if AI systems are capable of certain actions the MODD determines that there are certain actions which ought to be undertaken and others which ought to be avoided. The mutual compatible representations suggest one part of the ownership conditions, namely that the human in question is capable of under-

standing under which conditions the AI system does what. From the perspective of the AI system, the condition requires that they can keep track of the limitations of the user to avoid over-reliance (which is perhaps easier said than done!). The capacity to intervene on the actions of an AI agent, as well as explicit ties to actions of the agent and human agents, allow us to tie actions of the agent to humans and makes it possible for them to be involved in some significant sense.

The following question now arises regarding the type of AI technology we are focusing on in this paper. If we want to maintain meaningful human control over PAIAs in such a way that we can avoid some or all of the troubles described in the introduction, can we apply these four properties as design requirements? Among other things, this would require an explicit MODD, which can be comprehended by the user, such that they have the capacity to intervene if and when mistakes are made. This is, however, not as clear-cut as it seems.

3 Applying MHC to PAIAs

So, can we—and if so, how can we—make use of the notion of meaningful human control to aid in our search for design requirements for responsible PAIAs? Before knowing whether this is the case, we need to delve deeper into what exactly a PAIA may be. Since this is still partly to be determined, what follows is speculative. Suppose that you are in some higher-up position at a large and brilliant multinational, or that you occupy some other leadership position in which you regularly have to delegate tasks to others. You may have a human personal assistant (a PA) who helps you with your work, managing your schedule, discerning which e-mails you should read or even corresponding on your behalf, buying plane tickets, arranging visas and accommodations, carrying out background research and summarizing the findings. Perhaps they even occasionally buy gifts for some of your co-workers if you have forgotten to do that. In short, they make sure you are on time and know what you have to know at a given point and have the information ready at hand for the present situation.

Now, it ought to be clear that not all PAs are equal in terms of how well they perform. For example, they could be incompetent. In such a case, they would perhaps be sloppy, make weird inferences, mismanage your schedule, continually buy the wrong tickets, forget to arrange the correct visa (or worse, get you the wrong one), and find you subpar accommodations. They spend far beyond the expected budget. In short, managing affairs and the mistakes they make means they make your life worse off, and on top of that, increases your workload. They do exactly what they are not supposed to do. And you are likely not going to trust them with the tasks you give them. To have such a PA would require a lot of due diligence, and micromanagement just to make sure that the errors are fixed before you have to deal with the consequences. The power of a good PA has to do, in part, with taking away **your** worries and doing the correct action for **you** at the correct time. Such a PA finds ingenious ways to make your work and life easier and more comfortable, and perhaps even your decisions more informed and accurate.

Table 1 Applying MHC properties to PAIAs

Original properties	Specified for PAIA
The human-AI system has an explicit moral operational design domain (MODD) and the AI agent adheres to the boundaries of this domain.	The user needs to be informed of the positive and negative duties in line with which the PAIA should be operating.
Human and AI agents have appropriate and mutually compatible representations of the human-AI system and its context.	The course of actions undertaken by the PAIA are comprehensible. The human in turn can inquire about said course.
The relevant humans and AI agents have ability and authority to control the system so that humans can act upon their responsibility.	The decision-making on the part of the PAIAs has to be such that the user (or someone else) can monitor/observe the decision-making and intervene if they deem it necessary to do so.
Actions of the AI agent are explicitly linked to human agents, who are aware of their responsibility.	Information about the given trajectory has to be conveyed to the human, with the consequences of that trajectory being clear as well.

The way new PAIAs are described is often in terms of making life easier. This is not surprising, of course. The companies developing these AI agents are also interested in selling and marketing their products and services, so it is only logical that they would not advertise an ineffective AI. Furthermore, who would want a personal assistant designed to carelessly do such tasks? Can we perhaps use Meaningful Human Control as a design requirement to avoid such ineffectiveness? If we try and apply those concrete properties of MHC to PAIAs we may get something along the lines of Table 1.

How are we supposed to introduce MHC as a design requirement, given the variance and variability of the proposed technology? As we previously mentioned, PAIAs are undertheorized. We are given a range of options to choose from regarding a level of personalization but also in terms of stakes and preferences profiles. All of these cause issues. Can we solve them with MHC?

3.1 Issues of Duty

The first problem that comes with trying to use MHC as a design requirement lies within the MODD. Such a MODD requires that we further detail what the explicit positive and negative duties are in a domain, but that, like the PAIA itself, is underdetermined as well. Who is deciding the size of the domain? What is the extent of the context? Who gets to decide what the positive and negative duties are? Delineating a MODD for a PAIA requires that we are able to understand under which conditions and contexts such a piece of technology is implemented. Or, we could design a MODD with a specific domain and context in mind, to which the PAIA ought to adhere. Either way, it requires us to resolve contextual issues.

Consider, for example, contracts. Greediness in automated negotiation can cause harm for the individual in the long term, but it can also cause problems between indi-

viduals. If there is a large set of greedy algorithms for negotiating contracts, we may end up with a subset of users who take the brunt of the costs (e.g. the total utility of all contracts may be lower). Yet, can we enforce behaviour that leads to more optimal aggregate outcomes? In automated negotiation, contracts are notoriously difficult to design once we want a system that also keeps social welfare in mind (Jennings et al., 2001; Sanchez-Anguix et al., 2021). It can be done, perhaps, but what happens if an agent working towards social welfare runs into a greedy agent in a negotiation? If the stakes are low—e.g., we are merely interested in playing around with the agent for fun—then it is also easier to prefer short-term gains (ergo being greedy) for one's owner as it simply matters less. In higher stakes, this may be less warranted. Envision, an agent in the throes of negotiating a job for you. If such agents are greedy, they may accept some job that pays well but has no long-term perspective. If we want to enforce some social welfare, then we could ground it in an explicit MODD, but this may also hamper use and effectiveness, as well as the capacity of carrying responsibility. If, for example, we consider social welfare in a generous manner, we may argue that some action was done in the name of the greater good rather than for the sake of our own benefit. We could thus wash our hands of the blame. At the same time, if they are only working for us, then they may be extremely greedy to the extent that we are all worse off, this is a classical problem in automated negotiation (Baarslag et al., 2017). Greedy algorithms may work, but they may equally well create a situation in which some individuals end up victimized.

On top of that, higher stakes may make us vulnerable to losing control entirely over personal information or data. Not only do we need to ask what happens to our personal information and the company behind it that has it, such use of an advanced AI assistant is a highly desirable way to manipulate and influence individuals. If we rely on them massively, then both hackers but also the companies that can make profits from these agents' being used may want to game these agents in such a way that they can extract value from users. The problem of PAIAs in this scenario is that we give up direct control in favour of efficiency and/or effectiveness. Yet, this also means we risk losing control entirely. Di Nucci calls this the control paradox (Di Nucci, 2020). The balance to strike is one about the willingness we may have in giving up our direct control in favour of general ease. Yet, that ease may cause liability and responsibility issues and may make us vulnerable to losing control in general.

Negative and positive duties that come along with different levels of stake may cause us to consider to what extent we can rely on them. The control paradox in this case is a suggestion to what extent we may or may not want to enlarge the domain of the MODD. If the PAIA has a lot of different duties (both positive and negative) that will end up being a very different beast compared to one which is only allowed to play around in low stakes environments. The MODD will inevitably change the decision to whom to trace and what to track. In high stake environments, we will need to track different things as opposed to lower stakes. Fewer details may be relevant, and users may be more likely to take the blame due to the fact that they willingly participate in something that has few consequences in the real world. On the tracing side of things, the MODD will likely end up playing a major factor in understanding when the machine made a contextual error (i.e. went over the line of his negative duties) and thus which party to trace to.

Generally, the MODD requires knowing what the positive and negative duties are to actually implement them. This itself requires that we would know the contextual features of PAIAs. So in what environment are they nestled, whom do they interact with, what do they interact with? All of these matter to the duties in question, but for that, we need information that is (currently) not available to us.

3.2 Issues of Idealisation

A PAIA will have a range of tasks which it will need to execute well. Designing a system so that it does what you want is not as easy as it seems, as it will require knowing what is relevant to the situation (Kuilman et al., 2024). Everyone makes mistakes, and the same can be said of PAs which happen to be non-human agents. In fact, the PA in question is likely going to make a variety of mistakes. Yet, the difference between a human PA and an PAIA, cannot be understated. We are going to assume that these agents are products of companies, with whom our relationship is not the same as an interpersonal relationship with another human being. Even though we tend to anthropomorphize our relationship to technology in general, we have reasons to reign in these tendencies (Bryson, 2010; Nyholm, 2020). The fact that these PAIAs are products or services (rather than persons) makes them invariably different, experientially as well as legally and in terms of expectations (Evans et al., 2023; Milano & Nyholm, 2024).

One key issue is that of representation. Should the system have an idealised representation of your wants and needs? Or is it merely a copy of your current desires (this is often called a digital twin). There is the possibility of informed preferences (Gabriel, 2020), in such a situation, the system does what I really want it to do. What this really entails is exactly the problem. It could be: what I revealed to be my preference, what I say is my preference, or what I would do if I were rational and informed. The problem is also that the machine has to translate these things into practice (Russell, 2019). This translation is not a given, and neither is it certain that it actually portrays us well.

An idealised representation will likely need to capture the preferences of a user if we were to extrapolate it and include more time and knowledge. This may lead, in turn, to "better" results, however this does come at the cost of comprehensibility. It may not be needed that the user fully understands why the system is making a choice, but it needs to be interpretable. In idealised preferences, however, it would also require appropriate trust. Otherwise, it may be difficult to know when the system is going off the rails (e.g., when it is making a mistake or infers irrelevant things.). Such mistakes may undermine the feeling of responsibility towards such an agent. Even though we may have bought and paid for the service, and accepted an end-user licence agreement, that doesn't mean we will have the capacity to carry responsibility.

Idealisation may make the actions undertaken less understandable, this results in one of two things: either we would require that the human can either inquire to the extent that it isn't incomprehensible anymore, or we have to accept that there is a gap between the user's understanding of the situation and that of the system. The former is unlikely because it would require a lot of time (unless validation of preferences is very different from its calculation, but that seems unlikely) which hampers

effectiveness. The other option dampens our ability to intervene. While we do not want to micromanage such an AI assistant, the tracking condition of MHC will vary depending on whether we keep an idealised representation or not. In order to have MHC over these AI agents, we need the capacity to intervene once it is relevant. In other words, once it goes wrong. When higher stakes are involved (such as possible financial losses or risks of personal injury) we may require different capacities to intervene, we may want checks and balances beforehand. Yet, this may become much more of a hassle than we would like. Furthermore, if such systems are able to infer much more about our preferences (e.g. due to past behaviour), we also may not know whether they are doing something we reasonably desire, which could cause trouble for our intervention. For example, we could accept an outcome because we think the PAIA knows better than us, even though it is truly wrong.

If such a line of thought is promising, then such a system may actively harm our choice architecture, which is the basis for what Thaler and Sunstein call nudging (Thaler & Sunstein, 2021). Theoretically, such machines can help us make better choices. If based on our beliefs, viewpoints, and arguments, those choices can help broaden our search space for optimal solutions. Yet, in effect, there is also the potential for evil by design (Nodder, 2013), illicit use (Schmidt & Engelen, 2020), or what Thaler calls "sludging" (Thaler, 2018). The harms of nudging vis-à-vis the possibility of autonomy have been discussed more in-depth elsewhere (Schmidt & Engelen, 2020). Yet, it is profoundly important for PAIAs to get this right. With their proposed extensive use, we really need to confront the question: Who knows what is in whose best interest?

Idealisation causes issues for autonomy because it creates a lack of comprehensibility. Even with proper and sufficient interaction, there remains the possibility that we are mediated and lured towards a particular answer. If we try to counteract that by using current desires, we may diminish the effectiveness of the system.

3.3 Issues of Personalisation

Idealisation, of course, need not only be based on the individual. Idealisation can also be built upon preferences of groups. It can thus also be a matter of personalisation. Gabriel et al. (2024), sketch a difference between personal, semi-personal, and impersonal PAIAs, by which they look at highly individualized versions of such technology and those striving towards the greater good. In comparison, in *Human Compatible*, Russell (2019) argues that we should define the general idea of AI as artificial agents that should be working for the good of all of us. Personalised PAIAs are those built only on your preferences, whereas the other end of the spectrum, we could have PAIAs that only work in favour of all of us. In the middle, we can think of a PAIA working for a (small) group of people.

Both tracing and tracking require that we know to which extent the technology is personalised. If it is highly personalised, we may argue that the user is more at fault, as they mould the actions more so than if it is less optimized for an individual. Faulty behaviour in unpersonalised technologies may be more or less the fault of the designer instead, who picked an unsuitable behavioural pattern for the situation at hand. With tracking, we also need to worry about the personalisation, as this will

alter the relevant details to take into account. For example, how will such a machine capture preference drifts? The mechanics of that will determine how preference profiles may change over time. In highly personalised systems, preference drift will be personalised but in impersonal systems that may be drifting in terms of a population or even be static. What is traced and tracked matters depending on personalisation.

Furthermore, personalisation may not be similar in terms of our preferences. For example, we could be in a situation where a PAIA is excluding plane tickets from your travels because a majority stakeholder (that being other people) outweighs your desire to be somewhere in little time. Such a piece of technology might be under meaningful control, but perhaps not under *our own* meaningful control. For MHC we merely require someone to be in control, not exactly the user, but where we end up putting control has a major influence on how we will design these systems. Which will in turn decide what we require to correctly figure out tracking and tracing.

Individuals will disagree on what level of personalisation is preferable and what the greater good may be, which returns us to the point of idealisation (what would informed preferences actually be?) and the MODD (What would these duties entail?). Now, on the surface, this may seem like just another value trade-off. Is this not just another way of postulating the conflicts between the plurality that comes along with values and their alignment? We could view the situations as though we were simply debating whether to prefer one value over another (Petersen, 2021; Sutrop, 2020). However, considering that these systems will likely have more application compared to the current LLMs and will do more tasks, this means that the kind of value conflicts will penetrate many more spheres of daily life than ever before. It transcends the question of value conflicts because we likely will lose the ability to detract from such PAIAs.

If our preferences about these systems differ between individuals, we would also need to address that. Our capacity to understand a representation or our capacity to intervene is certainly not equal for all. A major problem for the implementation of technologies like PAIAs, is not merely because the technology is ill-defined, but is also in no small part problematic because we differ as individuals. This is technology for somebody, and that somebody is definitely not always the same person nor are they embedded in the same context. We are simply not all made equally in terms of our preferences and our use of such systems is likely dependent on our character, temperament, cultural background, intelligence, knowledge, and environment. So what access and for whom such systems may be, is highly dependent on who we are, and that causes an additional problem in design. It is not just a value conflict, with the purported extensive use, and even adjacent use (other people using it may inevitably impact our way of life as well), instead we run the risk of throwing up barriers for people. Before delving into that, there is one more thing we need to discuss.

3.4 Compounding Issues

We have discussed issues of duty, idealisation, and personalisation. Yet, all of these are linked. As mentioned, it is likely the case that negative and positive duties may alter based on the stakes, as well as the way that we conceive of its decision process. For example, if such a piece of technology works based on a highly informed prefer-

ence profile of the user rather than their current drives, it may be required to explain to the user why a certain choice was made. Otherwise, they would not have the capacity to intervene in a meaningful fashion. Yet at the same time, this may not be possible due to the number of choices made. If, rather than informed preferences, the PAIA acts solely based on a current preference profile, then the requirement for information might be lower—as the behaviour might be more insightful—and this could create a difference in requirements for explainability.

Delineating these issues makes it easier to comprehend them, but they do compound. Different preference profiles change the requirements of duties, as do stakes, as do personalisation. A high stakes environment with an idealized AI assistant might make choices and contracts which you as an individual may not want, even though they could be good for you. Idealisation and personalisation are also conjoined. A highly impersonal system may have different capacities for idealisation than one that is purely personalised. Again, the example of plane tickets would work well. Personalisation and idealisation work on different angles of the same topic, namely decision-making, but they do interact. An idealized preference profile of yourself will be different in a personalised situation as opposed to an impersonal situation. The idealisation would basically be different based on whom to take into account. But still, a digital twin that acts impersonally on your behalf is likely different from one that can take into account more time and knowledge.

Views on agency also play a contributing factor with all these issues. If we look at PAIAs through say a more agent-modelling approach (Bradshaw, 1997; Shoham, 1997), then we see that the relationship between PAIAs and agent approaches has in part to do with the proposed level of autonomy, reactivity, proactivity, and sociality. Yet, this means their analysis also has to happen more in the world (Yu, 2001) as their actions are far more dependent on worldly circumstances. The fact remains that PAIAs have to relate to worldly things, as well as being bound to epistemic limits of what can be modelled about the world. Considering that not everything in the world can be modelled, these issues described above are going to be paramount when thinking about what exactly to model and what an agent will be able to interact with in terms of domain and stake.

Nonetheless, all of these issues differ between individual users. Which context applies? What preferences do we have? What should be the size of the domain? And furthermore, how do we take into account the fallout of use for adjacent individuals? If we join these issues together, we may need to account for the fact that not all users are equal. They may have different informational requirements, and may require more explanation of the consequences that a certain trajectory has. This, in turn, may open these users up to more detrimental manipulation or require more trust as opposed to other users.

4 Enhancement and Meaningful Control

As we saw, there are problems with providing design requirements because of undertheorization. As a result, can we still say that this kind of technology provides a kind of enhancement? As Google's CEO mentioned, this kind of technology: *may*

give users a leg up. This idea of a leg up reminds one of the debates surrounding Human Enhancement and how far our interference can go in adapting human lives and improving our capacities (Bostrom & Roache, 2008; Juengst & Moseley, 2015; O'Neill et al., 2022; Savulescu, 2009). To what extent should we improve ourselves? The obvious historical approach is one best relegated to the past, but was about improving the human gene pool. Yet, the debate on human enhancement is also about who is going to benefit (Sparrow, 2016). It may even end up with human obsolescence (Danaher, 2018), in which this enhancement can interfere with the skills others have taken years to learn. Copywriting comes to mind in the recent age. And given certain societal rat races, it may alternatively mean that control is lost, simply because one has to keep up with the digitally savvy Joneses. We should not underestimate societal pressures and their relation to control. One can be pressured into using technology, even if one would rather not. The question is whether MHC can combat this, or whether this is outside the scope of MHC entirely.

Ideally, we create technologies to solve a problem or to improve our lives. Like freedom, technology can be a very good horse to ride, but we do need to ride it somewhere. It is this drive towards something which is currently lacking with technology like PAIAs. As Gabriel et al. (2024) mention, such systems may have a profound impact on our lives. They do discuss a variety of topics to show as much. For example, they discuss value alignment, and the potential for people with higher socioeconomic status to derive more help from such assistants. They are also aware of the social impacts and the sociotechnical systems that may arise from the introduction. But the question remains: To what end?

The relation between the "end" and MHC is essential if we want to understand the problems discussed thus far. If we want to track relevant moral conditions, we do need to know what we need to track. Otherwise, we may end up aligning with the wrong reasons. If technology has to enhance us, it also has to align to our reasons and to know those reason we do require contextual features in order to model those reasons.

With all of this in mind, we need to distinguish the problem of a leg up in two ways. Is it an enhancement for all? And is this taking us in a direction we desire? While the two are joined in some regard, it is helpful to separate these issues. It can be the case that such a piece of technology is more helpful to some than it is to others, or that it may even create a barrier. Moreover, while it may be helpful, it can push you in a direction you may not want to go. The benefit of talking about enhancement is that it presses us exactly on the question: What do we want from such technology and why?

4.1 Ubiquitous Enhancement?

Can this technology be ubiquitously good for all? Can we create technologies that enhance everyone in society? In this context, it is worth noting that Russell suggests in *Human Compatible* that we should define AI as technologies that solve all our problems. Yet, he realizes that we will have clashing preferences (Liao, 2020; Russell, 2019), so how should we understand this? Well, this starts with the obvious group that will have different values, those who are avidly opposed to the oncoming

of such technologies in any way shape or form. There have been staunch defenders against versions of technology, so with their inclusion, it is already not an enhancement for all. Gabriel et al. (2024) mention the values and goals of three different stakeholders: users, developers, and society. It should be clear that there may already be different incentives and conflicts between these three groups. We merely need to point back to the greediness mentioned 3.1 to see that social welfare can be opposed to certain gains for individuals. Enhancement, for all groups involved, is simply not immediately obvious.

Moving beyond these groups, we need to wonder what it means for such a technology to be good for all. Does it mean it has to improve everyone's life equally? Or merely that it improves everyone's life? If such an amendment to our personal lives is unequally distributed, it may not be seen as an enhancement. Consider, for example, a PAIA service that is offered by some tech company, which works for multiple people. What if these PAIAs do not work as well for you as they do for others? It could be that such a piece of technology is not only variable in what it can do, but also for whom it can do it and to what extent it can do these tasks well. Does MHC not imply something about the measure of control we have, or should have, over technologies? That also seems to imply something about the influence we can exert over technologies that we interact with or that make decisions for us (and about us). If formulated this way, then MHC still makes sense even in this larger societal context. In terms of the tracking condition, the PAIAs may track some users' reasons but clash with other users' reasons - and so a conflict arises about who should have MHC over the PAIA service(s) offered.

The reason why these systems may differ in effectiveness has much to do with the issues posed. Think about the differences between people. It could very well be that you have higher standards or that you request different things of technology. This would require that the MODD is particular to you, as opposed to what the average might be. Considering the potential personalised nature of PAIAs. We may simply run into the issues that non-personal PAIAs mean we are at a loss of control, since they act for us in a particular manner, but the same can be the case for personalised PAIAs. It is heavily dependent on what reason(s) are tracked and why. For example, if I am a highly private person compared to you, I may not be willing to give as much of my data away. Thus, if we are all forced into the ecosystem of using such technology, this may not be an equivalent upgrade for all.

There are also issues of idealisation. Consider, for example, if one has beliefs which cannot be represented by a machine (e.g. very holistic beliefs) This would mean that comparable representation does not correspond as well as it does for another individual. Does this mean that simply because one's beliefs, one's use of PAIAs is incomparable to that of your neighbours? Or you may belong to a group who is embedded in a particular context, age, or something else which hinders the PAIA from doing good work for you. It could even be a way for us to systematically deprive a group of people. An obvious example may be privacy conscious individuals, who for belief or other reasons are not willing or allowed to share their information (cf. Gabriel et al., 2024, section 13). Perhaps you don't grasp the full capacity of the machine and thus are unable to use it well. Regarding MHC, such a lack of under-

standing may also change our capacity to intervene, does that relate to our capacity to carry responsibility?

Other environmental factors may also impact the effectiveness. We could, for example, envision such technology interacting with other pieces of technology (IoT). If my smartwatch has an interaction with my PAIA, it may use some biodata as well to adapt one's preferences on the fly. Yet, I may not want those kinds of technologies in the first place. Another factor is whether one has a natural inclination to resist authority, a desire for more direct control, or, as we mentioned before, the desire for privacy. All of these impact both the capacity for ownership over the technology and the capacity for having the right representation. All of this in turn may affect the kind of stakes we should allow these agents to play with.

What ought to be adamantly clear is that such technology will differ in effectiveness among different individuals and groups. The simple way of understanding this is in terms of information. The digital divide may grow as the complexity of technology grows too, this means that having the capacity to wield such implements and understand artefacts may be important to get to parts of society. The negative framing of this is that we are in a process of both creating effective tools for some and effective barriers for others. The embedding of MHC requires multiple trade-offs that make it hard as a design requirement, but that stems from the fact that such PAIAs are undertheorized. We can thus also view these investigations as a sort of jumping off board, to understand what problems we might run into and what situations we may want to avoid.

If we desire ubiquitous enhancement, we would thus need to proffer different kinds of personalisation tools to different people, and amend their ways on different levels. For example, if we want it to benefit everyone equally, in what way do we benefit? If it is access to the same tool, we may inadvertently help those with more social capital and time to learn to use these tools as opposed to those who don't or can't. If it should benefit everyone equally, we may need separate tools for particular individuals to improve their lives more so than that of others. Both options seem to carry with them certain benefits and certain burdens that are hard to concretize without making choices about what to prioritize. Nonetheless, if we accept that these systems can profoundly impact us, such choices may also be seen as profoundly meaningful.

We need to grapple with the fact that this enhancement could cause a widening of the digital divide. An alternative may be to impose limits on those with additional skills or access or time, or even to provide additional benefits to those with fewer capacities and worse access. At the same time, skewing it that way may be unfair and it may also mean such extensive use allows for a heightened potential for error, and a greater loss of control. Besides, it is likely that those with time and money would want the best access. Putting a limit on them would be counterintuitive to the kind of market economies we have, and would likely require heavy regulation in the first place.

If we want to avoid inequality of enhancement, we require a delineation between those who fit some preconfigured mold and those who don't. That in and of itself can pose a serious threat to meaningful control because it would likely mean that certain things will be easier to control for some than for others. A standard for MHC in such technology may thus have the opposite effect, namely a loss of control, for some indi-

viduals. We could go further and even include cultural differences between individualistic cultures and more group oriented cultures, which may impact the effectiveness of PAIAs for entire cultures as well.

4.2 Directional Enhancement

Enhancement will be very ineffective if it is bringing us to the wrong place. That being, it enhances something that we don't want at the cost of a thing we cherish. If we want meaningful control, that also means that the trajectory is actually the one we want. Consider the problem in the aid of planning. We could have a PAIA which aids us in the creation of a perfect planning on a 10-minute basis. This could be highly efficient and effective, but it may, generally speaking, not be what we want. Enhancement in this sense is directional.

The previously mentioned under-theorization matters immensely because we may not all want to apply these systems in similar situations, nor do we want others to apply them in certain capacities. This is the reason why these issues are interlinked. We interact with others and if the standard of interaction changes (say less personal due to a heightened use of assistants) we are invariably affected even if we aren't the ones using said assistants. Yet, with regard to duty, idealisation, personalisation, we need to account for the fact that in which situations an assistant is applied and what it is allowed to do may make our lives better or worse. Some of us may not want to use it for the generation of ideas or content because we like to do it ourselves. Some of us may not enjoy outsourcing mental work because it may lead to a version of cognitive degeneration. At the same time, we might want to outsource certain tasks that we do not find enjoyable. What those tasks may be, will differ between individuals.

So what counts as enhancement? (Cf. Kahane & Savulescu, 2015) If one enjoys washing their clothes by hand, is a washing machine an enhancement? Well, it could be, if the surrounding sociotechnical system doesn't punish those who enjoy washing. If they are given ample time to do it in the way they enjoy and their life isn't altered in a meaningful fashion, it may be seen as a general addition to their life. If, however, there is societal or economical pressure to do something else, we may call this out as diminishment, at least for that individual. They may stop doing it out of pragmatic reasons because it is too time-consuming, even if they would prefer doing so it would carry with it certain costs because we have invented something new.

Now, should one individual's detriment stop the entirety of such an enterprise? Not necessarily, but we should take it into account, especially if it is not entirely clear what we are enhancing. Even if people do not enjoy a particular action, and we make it easier to do or access for many, it remains a question of what we are enhancing (and thus also for whom we are doing it). It does not seem unreasonable to say that certain unappealing activities may have benefits to individuals. In other words, the status of what actually counts as enhancement is not immediately obvious, which is why it helps to know what we are altering with technology and why we are doing it. This way we can help make better guesses who it may aid and who it might hinder.

The argument above is not meant to merely protect the status quo (Kahane & Savulescu, 2015; Savulescu & Kahane, 2009). Rather, if we devise an alternative through technology, we run the risk of creating a disability through social means. If PAIAs end up enhancing our cognitive capacities, then we are also basically creating a disadvantage for those who reject their use. And as Savulescu and Kahane point out, if we want to achieve enhancement, we do need some reasonable opinions on difficult topics like well-being and the good life.

The issues of duty, idealisation, and personalisation matter because the way we view context, how informed or revealed preferences affect us, may not be similar to all. But more importantly, the amount of personalisation may be essentially contested. It is debatable whether we will see eye to eye regarding personalisation, as this hinges on a notion of freedom weighed against a more communal notion. These differences are not only important in terms of barriers, but also what kind of society and interactions they end up shaping. As technology is a mediating force, we need to consider whether said mediation creates a desirable society. If we simply argue that this is an enhancement period, we do not acknowledge that it may be a trajectory we do not all agree with.

5 Conclusion

It is likely that the new AI assistants of tomorrow will have broad and general capacities. Currently, however, we've seen that such personal artificial intelligent assistants (PAIAs) are under-theorized. Which leaves many important aspects of design and deployment open to interpretation. If we desire such assistants to be feasible, we will need to consider additional design requirements. What we've provided here is one such attempt. A kind of jumping off board in hopes of providing new ways to look at such technology. We consider that PAIAs will need to take into account different possibilities of both stakes and stakeholders, as well as preference profiles of individuals, and the contexts in which they may be nestled. The problem that comes from under-theorization, however, is that it is very hard to know how to build design requirements. Our application of meaningful human control still has too many open-ended questions. We may not know how to operationalize a MODD, we may not know to what extent we should include idealised preferences, nor do we know to what extent we need to personalize it. All of these choices will have major consequences on how such systems will function and how well they will function for certain individuals. Even though these systems are potentially an enhancement for the user, as long as they are aligned with the desired *direction* of the enhancement in terms of what one wants. We also need to know what they are meant for. We need to make sense of what is being enhanced and at what cost, and who is benefiting of those - while remaining aware that achieving *ubiquitous enhancement* for all groups involved might not be feasible. Otherwise, it may be likely that this kind of enhancement is a barrier to some, and it may lead us to places that we do not desire to go in the first place.

Acknowledgements No acknowledgements to mention.

Author Contributions Sietze Kai Kuilman and Sven Nyholm contributed to the study conception and design. The first draft was written by Sietze Kai Kuilman and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding This research was supported by TAILOR, a project funded by EU Horizon 2020 research and innovation program under GA No. 952215.

Data Availability Not applicable

Declarations

Competing interests The authors have no relevant financial or non-financial interests to disclose.

References

Baarslag, T., Kaisers, M., Gerding, E., Jonker, C. M., & Gratch, J. (2017). When will negotiation agents be able to represent us? The challenges and opportunities for autonomous negotiators. In *International Joint Conferences on Artificial Intelligence*.

Bostrom, N., & Roache, R. (2008). Ethical issues in human enhancement. *New Waves in Applied Ethics*, 120–152.

Bradshaw, J. M. (1997). An introduction to software agents. *Software Agents, 4*, 3–46.

Bryson, J. J. (2010). Robots should be slaves. In *Close engagements with artificial companions: Key social, psychological, ethical and design issues* (pp. 63–74). John Benjamins Publishing Company.

Chamberlain, H. (2024). *The future of Ai with Google Ceo*. Youtube https://www.youtube.com/watch?v=h3M4bm2EveM

Danaher, J. (2018). Toward an ethics of AI assistants: An initial framework. *Philosophy & Technology, 31*(4), 629–653.

Danaher, J., & Sætra, H. S. (2023). Mechanisms of techno-moral change: A taxonomy and overview. *Ethical Theory and Moral Practice, 26*(5), 763–784.

de Santoni Sio, F., & Mecacci, G. (2021). Four responsibility gaps with artificial intelligence: Why they matter and how to address them. *Philosophy and Technology, 34*(4), 1057–1084.

Di Nucci, E. (2020). *The control paradox: From Ai to populism*. Rowman & Littlefield.

Evans, K. D., Robbins, S. A., & Bryson, J. J. (2023). Do we collaborate with what we design? *Topics In Cognitive Science*. https://doi.org/10.1111/tops.12682

Fischer, J. M., & Ravizza, M. (1998). *Responsibility and control: A theory of moral responsibility*. Cambridge university press.

Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds and Machines, 30*(3), 411–437.

Gabriel, Iason, Arianna Manzini, Geoff Keeling, Lisa Anne Hendricks, Verena Rieser, Hasan Iqbal, Nenad Tomašev, et al. 2024. "The ethics of advanced Ai assistants." arXiv Preprint arXiv:2404.16244.

Jennings, N. R., Faratin, P., Lomuscio, A. R., Parsons, S., Sierra, C., & Wooldridge, M. (2001). Automated negotiation: Prospects, methods and challenges. *Group Decision and Negotiation, 10*(2), 199–215.

Juengst, Eric, and Daniel Moseley. 2015. "Human enhancement." https://plato.stanford.edu/entries/enhancement/.
Kahane, G., & Savulescu, J. (2015). Normal human variation: Refocussing the enhancement debate. *Bioethics, 29*(2), 133–143.
Kudina, O. (2021). 'Alexa, who am I?': Voice assistants and hermeneutic lemniscate as the technologically mediated sense-making. *Human Studies, 44*(2), 233–253.
Kuilman, S. K., Siebert, L. C., Buijsman, S., & Jonker, C. M. (2024). How to gain control and influence algorithms: Contesting Ai to find relevant reasons. *AI and Ethics*. https://doi.org/10.1007/s43681-024-00500-y
Lara, F., & Deckers, J. (2020). Artificial intelligence as a socratic assistant for moral enhancement. *Neuroethics, 13*(3), 275–287.
Liao, S. M. (Ed.). (2020). *Ethics of artificial intelligence*. Oxford Academic https://doi.org/10.1093/oso/9780190905033.001.0001, accessed 9 Mar. 2025
Mecacci, G., Amoroso, D., Cavalcante Siebert, L., Abbink, D., van den Hoven, J., & Santoni de Sio, F. (Eds.). (2024). *Research handbook on meaningful human control of artificial intelligence systems*. Edward Elgar Publishing. Retrieved Mar 11, 2025, from. https://doi.org/10.4337/9781802204131
Milano, S., & Nyholm, S. (2024). Advanced Ai assistants that act on our behalf may not be ethically or legally feasible. *Nature Machine Intelligence*, 1–2.
Mitchell, M., Ghosh, A., Luccioni, A. S., & Pistilli, G. (2025). Fully autonomous ai agents should not be developed. arXiv preprint arXiv:2502.02649.
Nodder, C. (2013). *Evil by design: Interaction design to lead us into temptation*. John Wiley & Sons.
Nyholm, S. (2020). *Humans and robots: Ethics, agency, and anthropomorphism*. Rowman & Littlefield Publishers.
Nyholm, S. (2023). Responsibility gaps, value alignment, and meaningful human control over artificial intelligence. In *Risk and Responsibility in Context* (pp. 191–213). Routledge.
O'Neill, E., Klincewicz, M., & Kemmer, M. (2022). Ethical issues with artificial ethics assistants. In *The oxford handbook of digital ethics* (p. C17). Oxford University Press.
Petersen, T. S. (2021). Ethical guidelines for the use of artificial intelligence and the challenges from value conflicts. *Etikk I Praksis-Nordic Journal of Applied Ethics, 1*, 25–40.
Pisoni, G., & Díaz-Rodríguez., N. (2023). Responsible and human centric Ai-based insurance advisors. *Information Processing & Management, 60*(3), 103273.
Roff, H. M., & Moyes, R. (2016). Meaningful human control, artificial intelligence and autonomous weapons. In *Briefing paper prepared for the informal meeting of experts on lethal au-tonomous weapons systems, un convention on certain conventional weapons*.
Russell, S. (2019). *Human compatible: AI and the problem of control*. Penguin Uk.
Sanchez-Anguix, V., Tunalı, O., Aydoğan, R., & Julian, V. (2021). Can social agents efficiently perform in automated negotiation? *Applied Sciences, 11*(13), Article 6022.
Santoni de Sio, F., & Van den Hoven, J. (2018). Meaningful human control over autonomous systems: A philosophical account. *Frontiers in Robotics and AI, 5*, 15.
Savulescu, J. (2009). *Human enhancement*. Oxford Univ. Press.
Savulescu, J., & Kahane, G. (2009). The moral obligation to create children with the best chance of the best life. *Bioethics, 23*(5), 274–290.
Schmidt, A. T., & Engelen, B. (2020). The ethics of nudging: An overview. *Philosophy Compass, 15*(4), Article e12658.
Shoham, Y. (1997). An overview of agent-oriented programming. *Software Agents, 4*, 271–290.
Siebert, C., Luciano, M. L., Lupetti, E. A., Beckers, N., Zgonnikov, A., Veluwenkamp, H., Abbink, D., et al. (2023). Meaningful human control: Actionable properties for Ai system development. *AI and Ethics, 3*(1), 241–255.
Sparrow, R. (2016). Human enhancement for whom? In *The ethics of human enhancement: : Understanding the debate* (pp. 127–142).
Sutrop, M. (2020). Challenges of aligning artificial intelligence with human values. *Acta Baltica Historiae Et Philosophiae Scientiarum, 8*(2), 54–72.
Thaler, R. H. (2018). Nudge, not sludge. *Science, 361*(6401), 431–431.
Thaler, R. H., & Sunstein, C. R. (2021). *Nudge: The final edition*. Penguin.
Tielman, M. L., Jonker, C. M., & van Riemsdijk, M. B. (2018). What should I do? Deriving norms from actions, values and context. In *MRC@ Ijcai* (pp. 35–40).
Umbrello, S. (2020). Meaningful human control over smart home systems: A value sensitive design approach. *Humana.Mente, 13*(37), 40–65.

Wiesinger, Julia, Patrick Marlow, and Vladimir Vuskovic. 2024. “Agents.” https://www.kaggle.com/whitepaper-agents

Yeung, L. K. C., Tam, C. S. Y., Lau, S. S. S., & Ko, M. M. (2023). *Living with Ai personal assistant: An ethical appraisal* (pp. 1–16). AI & SOCIETY.

Yu, E. (2001). Agent-oriented modelling: Software versus the world. In *International workshop on agent-oriented software engineering* (pp. 206–225). Springer Berlin Heidelberg.

Authors and Affiliations

Sietze Kai Kuilman[1] · Sven Nyholm[2] · Stefan Buijsman[1] · Luciano Cavalcante Siebert[1]

✉ Sietze Kai Kuilman
S.k.kuilman@proton.me

Sven Nyholm
S.Nyholm@lrz.uni-muenchen.de

Stefan Buijsman
S.N.R.Buijsman@tudelft.nl

Luciano Cavalcante Siebert
L.CavalcanteSiebert@tudelft.nl

[1] Delft University of Technology, Delft, Netherlands

[2] Ludwig-Maximilians-Universität München, Münich, Germany