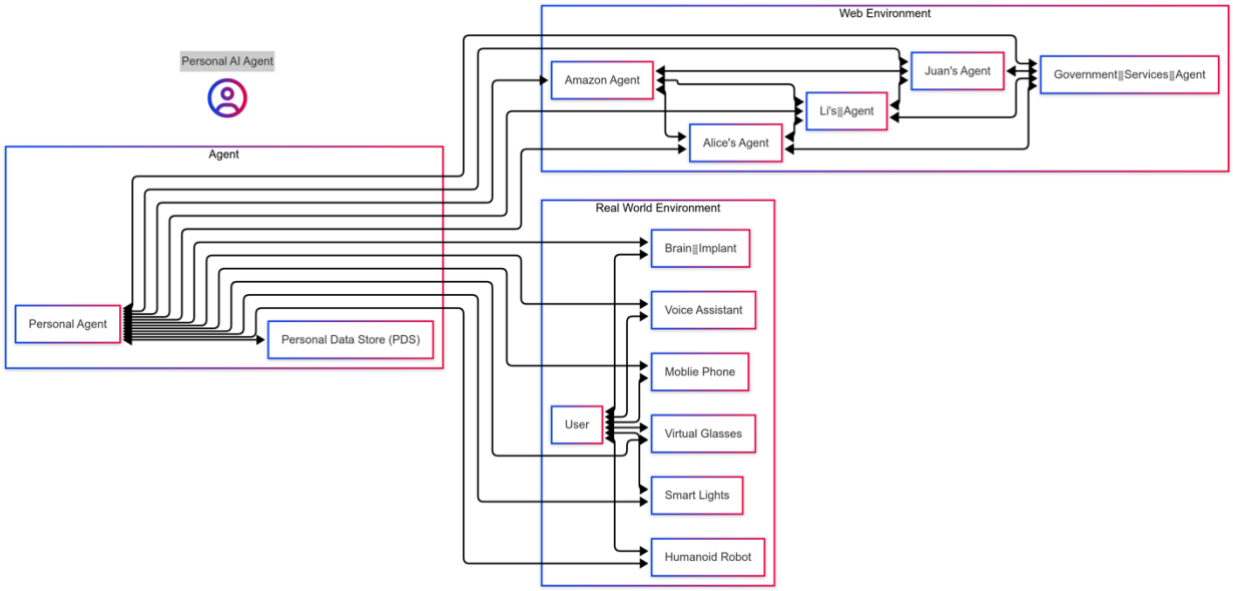
What do we mean by Personal AI Agents?

The notion of a Personal AI Assistant is not new. (Wooldridge, 2009) gives an example of a Personal Digital Assistant (PDA) which "converses with several different Web sites, which sell services such as flights, hotel rooms, and hire cars. After hard negotiation on your behalf with a range of sites, your PDA presents you with a package holiday." (Saad *et al.*, 2016) use the term Virtual Personal Assistant to describe "any **device** […] that provides professional, technical, or social assistance to automate or simplify daily tasks", and (Searls, 2012) use the term Vendor Relationship Management to describe the "customer-side counterpart of CRM, or customer relationship management […] that would make individuals both independent of vendors, and better able to engage with them."

The concept of AI existed as early as 1955, first coined by American Computer Scientist John McCarthy the focus was to "find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves" (McCarthy *et al.*, 1955). Today the scope and definition of AI is largely undefined. We view an 'AI system' to be a "machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments" (EU AI Act, Article 3). This means that a wide range of systems are idn scope for discussion, from non-interpretable 'black-box' systems such as generative Large Language Models (LLMs), geometric deep learning networks, and logistic regression classifiers which 'learn' to generate or predict outputs based on masses of sample training data, through to interpretable and predictable rules-based systems which execute a fixed set of instructions explicitly set by humans.

AI agents



So, what distinguishes AI and an AI agent? An agent is “a computer system that is situated in some environment, and that is capable of autonomous action in this environment in order to meet its design objectives” (Wooldridge, 2009) where autonomous action is the capability of agents “deciding for themselves what they need to do in order to satisfy their design objectives” (Wooldridge, 2009). Further, we expect the agents to be intelligent agents, characterized with **reactivity** – being able to understand, and effectively respond to their environment, **proactivity** – taking the initiative to service users, and **social ability** – being able to interact with the human they represent, as well as other agents.

What can constitute the environment for an agent is quite broad. For the purposes of this paper, there are two important environments in question – first, is the set of other agents that an agent will interact with, by messaging them on the Web. As we shall detail later in this section these agents will typically be either “Service Provider AI Agents” or other “Personal AI Agents” that represent different people. This is a typical multi-agent system (MAS) construction where agents communicate “not simply by exchanging data, but by engaging in analogues of the kind of social activity that we all engage in every day of our lives: *cooperation*, *coordination*, *negotiation*, and the like” (Wooldridge, 2009). By Wooldridge’s characterisation of agentic environments, this environment is *inaccessible* – as the agent does not have access to complete information about the action space or knowledge of other agents; *non-deterministic* as there are no guarantees as to how other agents are built – and hence respond, *dynamic* as the members of the agent system can change over time, and *continuous* as this multi-agent system is responsible for handling ongoing service interactions. Searls suggests that Personal AI Agents should primarily perform *intent casting* in this environment – for example, broadcasting the message “I want to buy 2 plane tickets from London to Berlin on Sunday Feb 9, 2025, departing between 6 and 9pm,” which Service Provider Agents representing airlines would bid serve. Intent can also be broadcast to other personal agents – for instance “I would like to meet with Janet on w/c Feb 12, please suggest times that would suit.”

The second environment we consider is the “real-world” environment in which the agent interacts with the user. This environment is the set of inputs provided by the user and their auxiliary devices, and the means by which the agent can respond or prompt. These auxiliary devices can range from an air-quality sensor providing data to the agent at fixed intervals, through to a voice assistant the agent can interact with, or a humanoid robot controlled by the agent. Additional user and auxiliary data may be made available to agents with access to Personal Data Stores such as a Solid Pod (Sambra *et al.*, 2016) – enabling agents to access any digital information collected about users, within the bounds of what users consent for the agent to access. By Wooldridge’s characterization of agentic environments, this environment is *inaccessible* – as the agent does not have access to complete information of the users’ world; *non-deterministic* due to the unpredictability of users and their environments, *dynamic* as users and their environment change over time, and *continuous* as the action space of the agent is not fixed nor is it finite.

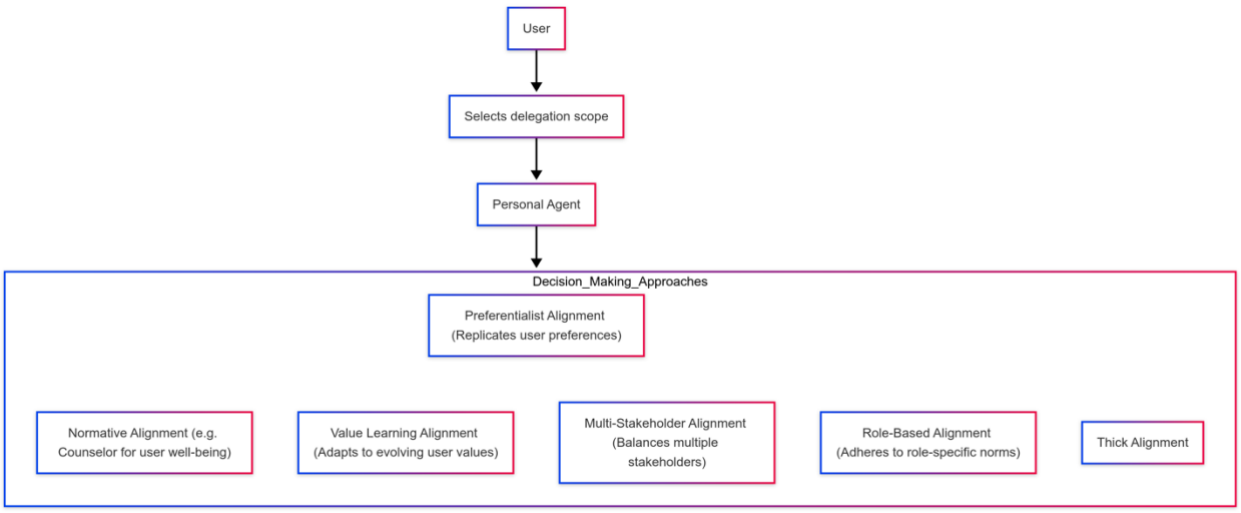
More recently, Gartner defined Agentic AI as "Autonomous AI can plan and take action to achieve goals set by the user" and have identified this as the top strategic trend for 2025 (Alvarez, 2024). However, Gartner envision these agents taking on normative roles (Zhi-Xuan *et al.*, 2024) within organizations – such as being integrated into a SaaS platform in order to replace some of the functions of a customer service representative. This is *not* in alignment with our vision of Personal AI Agents are *decoupled* from service providers and strictly represent the "best interests" of the consumer.

How consumers interact with agents

There are various modalities by which these AI systems may receive input and produce output. At first, there were algorithms called on demand by programmers running commands on their machine. This has significantly evolved over the last four to five decades, with the emergence of chat-like interfaces to interact with LLMs such as ChatGPT in 2021/2022, and now a rise in popularity of **Embodied AI (EAI)**. EAIs are AI systems with some form of physical embodiment – be it a webcam and screen providing a visual interface, or a full humanoid system with sensors that can capture all five human senses of sight (vision), sound (hearing), smell (olfaction), taste (gustation), and touch (tactile perception).

We consider both Embodied and non-Embodied Personal AI Agents to be within scope. We expect that just as with the AI services we interact with today, the modality with which it is appropriate to interact with an agent will be highly circumstantial. For instance, when consenting to having an agent purchase one's weekly shopping may take the form of hitting accept on a mobile notification; while planning a trip may involve a verbal discussion with a voice assistant to illicit preferences and allow a range of decisions to be made in a short period of time – much like working with a human travel agent. This multi-modality is a crucial feature when building personal AI Agents for vulnerable individuals.

Delegated control and decision making



When it comes to non-interpretable AI systems such as LLMs, there is increasing discussion around the topic of alignment. The traditional *preferentist* approach to alignment seeks to have AI systems understand the preferences of one, or more, users of the system and act in line with these preferences (Zhi-Xuan *et al.*, 2024). In cases where personal AI Agents have delegated authority and decision-making power (South *et al.*, 2025) this means that alignment results in a best-effort approach to emulate the decisions that the user would have made. More deterministic and rules-based systems is implemented by having users explicitly define what tasks an agent can *autonomously* perform; and the decision criteria that should be used when performing the task. A naïve instance of such an agent would be an email filter, which has a fixed set of rules to determine in which folder an email should be placed based on the sender and content of the subject. The kind of personal AI agents that are the focus of this paper, the constraints of what an agent is authorized to perform may be rules such as "do not spend more than $100 over the course of a week without my [the user's] authorisation," and the decision making criteria would be largely outline fixed preferences within particular task-scopes "when booking travel pick the cheapest hotel listed on my approved companies travel list, within a 500m radius of the conference."

For more contemporary machine-learning systems, a range of approaches are applied to align decision making preferences. One such emulative approach includes task-specific predictive systems – for instance, a machine learning system that identifies the products a user would buy by collecting a dataset describing the browsing history of a range of users, and the purchases they made – and then training a machine learning model to predict purchases based on user interaction with the browser over time. Note that this is the kind of predictive machine learning that powers targeted advertising in online platforms.

Similarly, the more generalist ChatGPT has been trained by "predicting" the sample output of a set of input text; and then having the response refined using Reinforcement Learning from Human Feedback (RLHF) such that the output is "defined by human judgment, building a model of reward by asking humans questions" (Christiano *et al.*, 2017; Ziegler *et al.*, 2019). In cases such as that of ChatGPT, this process of RLHF is *not* done to align the system to a set of individual user preferences; instead, the system is being trained to comply with specific normative criteria (Zhi-Xuan *et al.*, 2024) including "helpfulness, harmlessness, and truthfulness" (Ouyang *et al.*, 2022; Bai *et al.*, 2022). These normative roles are communicated to the human employees and contractors of OpenAI tasked with providing the system feedback for RLHF.

In both the predictive-purchasing, and ChatGPT example; these systems are *not* being designed to emulate the preferences of an *individual user* but rather be predictive of the behaviors of a population at-large. In contrast, we expect that if a personal AI Agent uses *machine learning* and is *preferentialist* then the system would specifically try to emulate the *user intent* when decisions are delegated to the agent.

This both calls into question how we align with user intent and whether we should be aligning with user intent at all. As to whether it is possible to *align with user intent*, Zhi-Xuan *et al.* observe that the traditional *preferentialist alignment* (Baber, 2011) approach for machine-learning AI systems makes the false assumption that *humans are themselves rational decision makers*, that can *capture their values as a set of preferences* and *always act to maximize those preferences*. When this assumption breaks; it becomes very difficult for a system to discern a clear set of criterion upon which to establish if it is following user intent – much as a human personal assistant can only roughly guess the decision making procedures of their superior, and never fully emulate them.

There is also a further discussion of whether we should be instead building systems that are not *value* or *preference* aligned, but instead "optimised" in other ways – such as making decisions that are in the interest of the users' long-term wellbeing. Zhi-Xuan *et al.* suggest that systems should always be designed to fulfil normative societal roles – such as a travel planner, psychologist or manager. Some argue that we should perform **thick value alignment** to ensure AI is aligned with human values at large (Russell 2019, 137). Ji *et al.* suggest that when doing such **thick alignment** there are four guiding design principles to be accounted for Robustness, Interpretability, Controllability, and Ethicality (RICE). We highlight this as a critical open ethical question in the design of personal AI Agents.

Noting all the above alignment challenges, we expect that in the near term, most Personal AI Agents will be a hybrid of deterministic rules-based systems and black-box symbolic systems – a simplistic example of this is presented in Wright (2025). For the most part, we expect that the user delegates control to the agent using rules-based "authorisation controls" (South *et al.*, 2025) and within these bounds a neurosymbolic system performs decision making according to some form of alignment.

Distinguishing personal and personalized AI

Personalized AI is characterized by being in some way tailored, or in some way self-tailoring for a particular user. Examples of Personalized AI Agents include *recommender systems* (Ko *et al.*, 2022), *smart home assistants* (Saad et al. 2016; Santos et al. 2016)*, conversational LLM's such as ChatGPT* and *Computer Using Agents (CUAs)* such as OpenAIs operator agent[1]. What most of these personalized agents have in common, is access to some degree of personal data with which to inform their interactions with users. For *recommender systems* it is previous watch history to prescribe suggested shows, *home assistants* have access to calendar data to alert you of upcoming events *Amazon Alexa* further supports contextualized discussions – such as about one's interests, and learns repeated user behaviors and notifying them with a "hunch" that they may have forgotten something. Another common feature is tailored mannerisms. Voice assistants such as *Amazon Alexa* which have customized voice profiles, *ChatGPT* - which uses past conversations with a user to provide context to the current

[1] https://openai.com/index/introducing-operator/

conversation, thus making the result more *relevant* to users; and also makes the conversational LLM begin to act *more like the user*[2].

While we expect all *Personal AI Agents* to be *Personalized AI* – the converse is rarely true. The earlier discussion around alignment is what fundamentally distinguishes the personal AI Agents we discuss in this paper from personalized AI, which is more prevalent in the existing service literature. Alexa is a good example where the system is not aligned with user *intent* or interests – as users are often recommended to buy products by the device; not because they are what the user would normally choose to buy, or are necessarily in their best interest to buy, but instead because the system is marketing a product to them. This is exactly why there is a need for *personal* agents which advocate for users.

Modern Personal AI Agents are beginning to emerge. Kwaai.ai[3] for instance is a non-profit lab building "a [self-sovereign] Personal AI Operating System to allow you to train your own personal assistant, privately [and] securely."[4] It is led by Doc Searls who invented the concept of the *intention economy* and *vendor relationship management.* To some extent open source frameworks such as BabyAgi[5] may also be considered to be working towards Personal AI Agents, by laying the groundwork for end-users to design custom AI agents for themselves.

We need to consider how we can ensure that Personal AI Agents are not operating with ulterior motives when deployed in practise – for instance, how can we know that our personal AI Agents are not just *Personalized AI Agents* in disguise and ultimately working in the interest of a particular organization by "manipulating" us to buy specific products or services, much as current "attention economy" services do today (Searls 2012). One approach is to encourage the development of open-source implementations of Personal AI Agents such as kwaai.ai, which can then be deployed locally on individuals' devices – while a nice ideal, this still requires most end-users to trust opensource developers in their design and implementation of such agents, with very little means to understand what has been done. A more compelling answer may lie in making companies that implement services for Personal AI Agents legally accountable – and subject to fines if the agents they implement are found to be in anyway make decisions to better the commercial interests of the company rather than the user once deployed. If too heavily regulated, however, this risks stifling the development of such agents.

What we mean by Personal AI Agents

To summarize a personal AI agent is an agent operating within a multi-agent system of personal AI agents, and service provider agents. Personal AI agents *must* represent user interests, that is, have **alignment** with the values and intentions of the individual user when

[2] https://help.openai.com/en/articles/8590148-memory-faq
[3] https://www.kwaai.ai
[4] https://drive.google.com/file/d/1lHxy0q3z5krG8hBwYIG6bDloIuzagSce/view
[5] https://babyagi.org

given authority to act autonomously, and support their self-determination (Ibáñez *et al.*, 2023). The scope, or granularity, at which agents may which act autonomously is to be user defined – if deployed at scale, we anticipate that there will be a range of preferences that users have for the degree of autonomy they wish to delegate; some customers we expect to place expectations such as "notify me before making any *social, legal* or *contractual agreement*," while others may set looser bounds for autonomy "bring me into the loop if you plan to spend more than £200 in the course of a week."

In contrast to most AI Agents, and Personalized AI the service literature, these agents *are not* to be implemented by providers of a particular service – but instead interact with the service provider while representing user interests.