Robust AI Personalization via The Human Context Protocol

Anand V. Shah
Massachusetts Institute of Technology
avshah@mit.edu

Tobin South
Massachusetts Institute of Technology
tsouth@mit.edu

Talfan Evans
Persona Labs
talfan@personalabs.co

Hannah Rose Kirk
University of Oxford, UK AI Security Institute
hannah.kirk@oii.ox.ac.uk

Andrew Trask
OpenMined
andrew@openmined.org

E. Glen Weyl
Microsoft Research
glenweyl@microsoft.com

Michiel A. Bakker
Massachusetts Institute of Technology
bakker@mit.edu

Abstract

This paper argues that robust AI personalization requires a Human Context Protocol (HCP): a user-owned, secure, and interoperable preference layer that grants individuals granular, revocable control over how their data steers AI systems. By replacing siloed, behavior-inferred signals with direct preference articulation, HCP unifies fragmented data, lowers switching costs, and enables seamless portability across AI services, fostering a more competitive ecosystem. We outline core design principles – natural-language preference storage, scoped sharing, and strong authentication with revocation – that extend earlier personal-data architectures to the scale and stakes of modern generative AI. Centering control in users, HCP is not merely a technical convenience but a necessary foundation for AI systems that are genuinely personal, interoperable, and aligned with diverse human values.

1 Introduction

Large language models (LLMs) are rapidly becoming embedded in everyday digital experiences, transforming how people access information and services. Central to unlocking their full potential is "personalized alignment" – tailoring model behavior to reflect individual preferences, values, and contexts [Kirk et al., 2024]. The evolution toward personalization is accelerating rapidly, with major AI providers including OpenAI, Google DeepMind, and Meta having announced personalization features in 2025 as central axes of their development roadmaps.[1]

However, the current paradigm for personalization presents significant challenges. Preference data, when captured, is typically inferred from behavioral traces that may poorly reflect true user intent, leading to shallow or inaccurate personalization [Kleinberg et al., 2024b]. Furthermore, this user data is often fragmented across services and locked into specific providers, reinforcing user lock-in [Farrell

[1] Respectively, this is OpenAI [2025], Google [2025], and Meta [2025].

Preprint.

and Klemperer, 2007]. This lack of interoperable, user-controlled infrastructure also undermines progress on AI alignment. In the absence of explicit, fine-grained user signals, models are often aligned via biased feedback from small, unrepresentative rater pools – what Kirk et al. [2024] term the "tyranny of the crowdworker."

Recent initiatives like the Model Context Protocol (MCP) aim to create open standards for connecting AI assistants to wide-ranging data sources [Anthropic, 2024]. While valuable for enabling context-aware AI by standardizing data access, MCP can not address critical questions of data ownership, granular user control, and privacy management for personal preferences. Yet, these questions are critical to the actual deployment of personalized AI solutions.

We argue that a dedicated user-centric architecture for preference management is a core requirement for building AI systems that are truly personal, interoperable, and aligned with diverse human values. We propose the "Human Context Protocol" (HCP): user-owned, secure repositories of preferences designed for active, reflective control and consent-based sharing. This infrastructure aims to give individuals agency over how their data informs model behavior, allowing user preferences to port seamlessly into LLM agents. Building on earlier work in personal data stores, our proposed architecture enables individuals to:

- Maintain ownership of where their preference data goes, granting fine-grained, revocable access to various LLM-powered services.
- Transfer preferences between AI providers much as mobile users may switch phones, reducing switching costs and provider lock-in.
- Actively participate in shaping how their preferences inform model behavior through natural language interfaces.

Moreover, we contend that HCP will not exist without dedicated effort from researchers, as market forces alone may not incentivize providers to relinquish their control of user data. This user-centric architecture addresses a critical gap: the absence of infrastructure allowing individuals to privately self-manage their preference data – an essential foundation for a user-empowered AI ecosystem.

2 Background and related work

The tension between privacy and personalization has driven successive waves of theory and product for personal-data control. Initial discussions on privacy centered on personal dignity and the right to self-disclosure [Warren and Brandeis, 1890, Westin, 1968]. Yet, as online data proliferated – often invisibly and at immense scale – this individual control was increasingly undermined, leading digital scholars to expand the frame of privacy to include protection from commercial exploitation [Laudon, 1996, Varian, 1996] and inspiring designers towards architectures that prioritize agency.

2.1 Work on personal data

The modern genealogy of user-controlled data begins with Hagel and Rayport's 'infomediaries,' imagined brokers that would negotiate data use on the individual's behalf [Hagel III and Rayport, 1997]. Although visionary, infomediaries never overcame the two-sided-market adoption barrier, requiring buy-in from both users and firms in a time where internet markets were still nascent.

A more durable ideological basis for personal data control emerges in movements like Europe's MyData, which articulated human-centric principles such as portability and individual data sovereignty [Poikola et al., 2015]. Tim Berners-Lee's Solid project operationalized similar ideals in "pods" – decentralized architectures where users store data and manage access via revocable permissions [Sambra et al., 2016].[2] The Self-Sovereign Identity (SSI) movement extended this logic to digital identity, arguing that identifiers should be user-controlled rather than issued or maintained by central authorities [Allen, 2016, Mühle et al., 2018]. More recent implementations (particularly Web3-enabled "data wallets") extend this model further, aiming to give users custodial control over identity, reputation, and other personal data using cryptographic methods [Zyskind et al., 2015]. While there

[2]Protocols like Solid offer a viable backbone for HCP infrastructure, providing robust decentralized data storage. This can be further augmented with tools such as MCP and local orchestration LLMs.

has been much work on building independent personal data stores, these efforts have yet to yield a widespread user-controlled preference management solution.

Recent advances in AI may change these dynamics in two material ways. First, the value proposition for users contributing preference data has increased substantially. User data now supports increasingly capable AI systems that function as general-purpose assistants, and preference data further personalizes AI systems to the user themselves [Christiano et al., 2017, Ziegler et al., 2019, Ouyang et al., 2022]. Second, the emergence of natural language as the primary interface modality for AI systems substantially reduces the cost of expressing and updating preferences. Textual input offers a more accessible and natural means for users to articulate complex contextual information and preferences. A comparison of HCP to previous artifacts of personal data control are summarized in Table 1.

Table 1: Evolution of User Data Control.

Initiative (Era)	**Key Idea**	**Mechanism**	**Limitations**
Infomediaries (Late 1990s)	Brokered user data via intermediaries	Third-party agents managing consent	Indirect control; user frictions; requires large market adoption
MyData (2010s)	Data sovereignty as a civic right	Normative principles	No technical practicality
Solid Project (Mid 2010s)	User-controlled decentralized storage	Data "pods" with revocable permissions	User frictions (self-hosting); ecosystem still developing; limited scale in natural language scoping
SSI (Mid 2010s)	Portable, user-owned digital identity	DIDs and verifiable credentials	Limited to identity attestations; architecturally unsuited for rich data
Web3 Data Wallets (Late 2010s)	Custodial control over digital assets	Keys, smart contracts, blockchain	Very high user frictions (DeFi management); no rule of law; on-chain-asset centric
HCP (2025)	**User-directed preference management**	**LLM-native preference interface**	**Adoption requires ecosystem buy-in; ensuring security & mediating LLM integrity is crucial.**

2.2 How personalization is done today

First, we make a distinction between *inferred* preferences – deduced from user behavior and interaction patterns – and *explicit* preferences – directly articulated by the user.

While AI providers rarely acquire outright ownership of user data (users typically retain the right to delete chat histories or opt out of future training), they often maintain broad, perpetual licenses to use, modify, and even sublicense user data, especially for the use of preference inference.[3] In this sense, providers can be said to retain substantial de facto property rights over user data, as their broad licenses grant them control over most residual claims regarding its use [Grossman and Hart, 1986]. This permission structure means AI personalization is dominated by the provider side, and also explains why future AI personalization may largely rely on inference methods in much the same way that today's digital personalization is undertaken. In the absence of inferred personalization, companies may also offer relatively coarse, global settings within their applications, such as OpenAI's "Custom Instructions" for ChatGPT, which allow users to provide explicit, high-level guidance.

Further, while AI systems can powerfully search and analyze past conversations to build memories of user interactions [Chhikara et al., 2025], memories differ fundamentally from preferences. Memories

[3]For instance, this is true across the privacy policies of OpenAI, Google Gemini, and Meta AI [OpenAI, 2024, Google, 2025, Meta, 2025]. In May 2025, Gemini launched a "personalized" model fine-tuned with an agent's preferences (learned from browsing history observed on Google, linked via email).

contain raw conversational data that may include preference information, but extracting meaningful preferences requires further inference. Unlike memories, preferences are explicitly stated values that users can directly edit and control. The security features we propose in Section 3 – such as fine-grained access controls and selective sharing – work naturally with preferences but become unwieldy when applied to the vast, unstructured space of memories.

Beyond directly personalizing models with inference or memory, significant progress in creating models amenable to personalization has been achieved via fine-tuning, using techniques like RLHF or DPO [Ouyang et al., 2022, Rafailov et al., 2023]. These generic fine-tuning strategies often aim to produce a pliant base model that can be more easily personalized by downstream users via prompting or few-shot examples – in this sense, preference data is inferred from the prompt itself, and models compete along their ability to provide the optimal output for a user context given any particular prompt. These general techniques can also be modified to more directly tune on specific signals (e.g., as in Poddar et al. [2024] or Li et al. [2024]).

Of course, a tuning algorithm is only as effective as the user data disciplining it. A core limitation of current systems is that preference data – whether inferred or explicitly provided – remains siloed within individual providers. Users lack transparency into how their data is used, cannot port data across services, and have minimal control over its evolution or erasure [Kirk et al., 2024]. This also creates a substantial "fragmentation tax": users must re-specify their preferences across agents and firms, leading to inconsistent experiences, increased cognitive load, and growing user frustration as personalization must be repeatedly rebuilt from scratch.

The problem of fragmentation becomes more acute as users interact with multiple AI systems across different aspects of their lives – work, entertainment, health, finance – each maintaining a siloed, partial view of the individual. This not only limits personalization fidelity but also forecloses potential, complementary cross-domain insights. Even as individual models improve, the user's experience across services becomes increasingly incoherent.

3 System design

To realize the vision of an HCP ecosystem, any system design should have the following core attributes:

- **Interoperable**: HCP must be interoperable across AI models and application contexts, as this is fundamental to its utility. Interoperability should be facilitated by open, well-supported, and existing communication protocols.
- **Sufficiently Representative**: For HCP to provide genuine utility, it must be capable of richly capturing user preferences. While current models have limitations in preference elicitation and representation, the HCP's data model should aim to be at least as expressive as state-of-the-art techniques in preference capture.
- **Scopable Sharing**: Given the personal and sensitive nature of preference data, users must have fine-grained, revocable, and editable control over what preference information is shared and with whom. For instance, a user should be able to share culinary preferences with a recipe generator without exposing mental health information.[4] This aligns with the principle of data minimization, ensuring only necessary information is disclosed for a given query.
- **Secure**: The storage and transmission of sensitive personal data within HCP demands robust security measures. Preferences must be secured at rest and in transit, with strong authentication and authorization mechanisms to ensure AI models only access explicitly authorized preference subsets [South et al., 2025].

3.1 Proposed solution

This paper does not prescribe a definitive implementation for HCP; any system fulfilling the aforementioned design attributes would be suitable. However, to facilitate discussion, we outline a *potential* protocol architecture below (illustrated in Figure 1).

[4]We believe revocability is an important part of scopable sharing. However, various agent models may interact with revocability in different ways – for example, if user data is used in fine-tuning, such as through RLHF, revoking preferences might necessitate reverting to a previous model checkpoint or retraining.

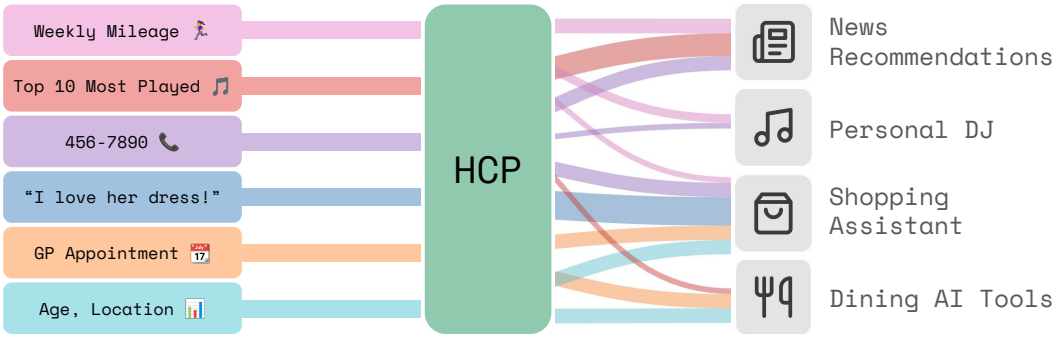

Figure 1: Illustration of HCP. User preference data (generated by varied user activity) is moderated by HCP to consumer agents. Each agent obtains only the relevant subset of the user's complete preference data. In the current ecosystem, agents themselves are built upstream on MCP.

3.1.1 Natural Language Representation

We propose a system where **preferences are stored and managed primarily in natural language**, potentially augmented with images. This format aligns with the dominant input modalities of most LLMs, providing more interoperability than custom formats or learned embeddings. While the expressive sufficiency of natural language for complex preferences remains an open question, existing functionalities – like ChatGPT's custom instructions or memory features from chat history as in [Chhikara et al., 2025] – suggest its viability. Mechanisms to mitigate potential ambiguities inherent in natural language, allow user clarification [Pyatkin et al., 2022], and manage the evolution or versioning of these expressed preferences will be important considerations.

3.1.2 Managing Selective Disclosure

A comprehensive, multi-modal natural language corpus could richly represent an individual's preferences, but **managing selective disclosure** from such a large dataset is challenging. Schema-based approaches – categorizing preferences into fields like 'food' or 'media' – enable fine-grained access control, yet introduce user experience concerns at scale. For multifaceted tasks like planning a dinner party, users may need to authorize access across many overlapping or unforeseen categories (e.g., dietary restrictions, preferred locations, social ties, past event feedback). Requiring meticulous permissioning for each interaction quickly becomes burdensome.

An alternative approach leverages an LLM as the core information management tool within the HCP itself. This *dedicated 'HCP LLM'* (potentially a smaller, locally hosted, or specialized model) would interpret access requests from external AI services and selectively provide only the relevant preference information at inference time, thereby directly implementing data minimization. The trustworthiness and careful alignment of this HCP LLM itself is paramount. This architecture separates the preference management function from the primary LLM interaction, offering several advantages: it enables robust logging of all shared preferences, provides a clear audit trail for user transparency, and allows the HCP LLM to be independently evaluated and hardened for its specific role.

3.1.3 Storage and Security Architecture

The underlying storage mechanism for preference data can be flexible. Options include a single comprehensive document, a key-value store with access permissions, a graph-based representation [Pan et al., 2024], or comprehensive solutions like Solid [Sambra et al., 2016]. For highly sensitive preferences, the HCP could support user-held encryption keys, with the HCP LLM prompting for decryption credentials at access time. For preference sets exceeding typical LLM context windows, vector databases enabling semantic search present a viable alternative (potentially with added privacy [Zyskind et al., 2024]).

Security considerations must address both standard practices (encryption at rest and in transit, authenticated access, and authorization protocols) and emerging threats specific to LLM systems. Future work should explore defenses against inference attacks on shared preference patterns, mechanisms for

verifying preference integrity, and methods to protect smaller HCP LLMs from adversarial prompting by more capable external models.

3.1.4 Protocol Interfaces and Interoperability

To concretize how HCP operates in practice, consider the protocol's key interfaces. These could manifest as RESTful endpoints (e.g., `GET /userInfo` for basic facts) or MCP tools providing similar functionality. The protocol supports both *informational* operations that retrieve static data and *interactive* operations enabling dynamic preference management. For instance, `addPreferences` and `searchPreferences` would function as read/write operations on a preference database, potentially augmented with RAG techniques for relevant information extraction.

The granularity of preference categorization represents a key design dimension. Implementations might range from unified endpoints consolidating all user data to fine-grained separation of distinct data types (e.g., `searchPreferences` versus `searchFoodPreferences`). This flexibility allows the protocol to adapt to diverse use cases while maintaining core functionality.

A Concrete Walkthrough. Consider a user integrating HCP with their Claude assistant. The user initiates this by adding the HCP-MCP server from their available integrations list, completing a standard OAuth 2.0 authentication flow. During a conversation about hair care, when Claude recommends a specific shampoo brand, the user requests: "add this to my preferences for later." This triggers a tool call where the relevant information is extracted and stored in the HCP's key-value store. Later, when the user inquires about hair products, the `searchPreferences` tool performs retrieval-augmented generation across stored preferences, returning the previously saved recommendation. The entire interaction occurs seamlessly while maintaining user control over data storage and access.

Connecting HCP to various AI interfaces can leverage emerging standards like MCP [Anthropic, 2024], Agent-to-Agent (A2A) frameworks [Google, 2025], or LLM tool-use functionalities. All access must be authenticated and explicitly authorized (e.g., via OAuth 2.0 [Jones et al., 2015]), with revocation remaining at the user's discretion.

3.2 Demonstrating feasibility: an open-source prototype

The proposed conceptual design is readily implementable, a crucial characteristic for fostering an open preference ecosystem. To demonstrate viability and encourage further work, we developed an open-source proof-of-concept.[5]

This prototype embodies several core attributes. It is a web application where users manage preferences and control third-party access. An integrated MCP server supports interoperability with compatible AI interfaces, enforcing user authentication and granular authorization for distinct preference categories. Preferences are stored as categorized textual data in PostgreSQL. When an external application queries the HCP, a dedicated model (here, GPT-4o serves as the HCP LLM) ingests only authorized preference categories and provides only the relevant information subset, demonstrating practical inference-time data minimization.

While an early step, this prototype confirms the feasibility of constructing an HCP that is interoperable, secure, and grants users meaningful data control. It provides a foundational codebase for the community to build upon.

3.3 Self-sovereignty and custodial control

The ideal of self-sovereign identity, where users have ultimate ownership and control over their digital assets without reliance on intermediaries, resonates strongly with HCP's ethos. This approach enhances user autonomy and privacy through user-controlled storage (self-custody) and minimizes trust in any single service provider. The benefits include greater resistance to censorship and provider lock-in.

However, implementing true self-sovereign storage and key management for all users introduces significant usability challenges and technical overhead, potentially hindering broad adoption. The core attributes of HCP – particularly fine-grained, scopable sharing mediated by a trusted component

[5] See *https://github.com/tobinsouth/hcp-demo*.

like an HCP LLM, coupled with strong encryption and clear authorization protocols – can provide substantial user agency and data protection even when users entrust their encrypted HCP to a third-party custodian. The critical factor is preserving user control over disclosure decisions. HCP is fully compatible with more decentralized, self-sovereign approaches and could evolve in that direction, but does not require self-custody as a precondition for achieving its primary goals: secure, interoperable, and user-directed preference sharing.

4 Arguments for HCP

4.1 The Coasean challenge: does architecture matter?

A naive application of economic intuition might suggest that it is inconsequential whether users or firms control preference data. Coase's theorem states that under zero transaction costs and well-defined property rights, a perfectly competitive market will achieve the efficient outcome regardless of the initial allocation of property rights [Coase, 1960]. In our setting, as long as preference data remains contractible, the market will equilibrate to the efficient outcome regardless of whether they're owned by users or firms.

While the discussion is not purely one of property rights,[6] Coasean reasoning might lead us to believe that the architectural locus of preference control – whether vested in users or firms – is not very relevant for social welfare. Yet, several factors drive a wedge between theory and reality.

4.2 Market failures

Below, we describe some particular market failures that may be resolved by HCP.

First, **lock-in and interoperability** constitutes a failure of the zero transaction cost assumption. When preference data is locked within specific service silos, users face high switching costs if they wish to employ a competing agent or service. This friction limits user choice and dampens competitive pressure on agent providers to improve quality or compete on price [Varian et al., 2004]. For example, prior to phone number portability regulations in telecommunications, switching carriers also meant losing one's number – a critical piece of digital identity. The introduction of number portability dramatically increased competition and reduced prices [Viard, 2007]. Similarly, HCP, designed with interoperability as a core principle, would reduce switching costs and foster a more dynamic ecosystem in which agents compete on performance.

Second, **the non-rival nature of (preference) data** constitutes a failure of property rights. Unlike physical goods, data is non-rival – its use by one entity does not diminish its availability for others.[7] This characteristic implies that social welfare is maximized when valuable data is used broadly, subject to privacy constraints. However, when firms control user data, competitive incentives lead to inefficient data hoarding; firms are reluctant to share data that might empower rivals or accelerate their own creative destruction [Jones and Tonetti, 2020]. HCP, by assigning control to the user, provides a mechanism to ameliorate this market failure – users can choose to license their preference data as broadly as is useful for themselves, enabling the aggregate productivity gains typically associated with information goods.

Third, **information asymmetries and market power** constitutes a departure from perfect competition on the firm side. Large firms often possess far more information about market conditions and user behavior than individual users do, along with the analytic tools to exploit that asymmetry. For example, Acquisti and Varian [2005] describe precisely this dynamic in data markets – firms extract user surplus by leveraging purchase history to conduct targeted pricing. By giving users control over the release of their preference history and associated information, HCP empowers consumers to strategically manage (exploitation from) their information footprint.

[6] See also the discussion in Section 2.2 on property rights.

[7] More precisely, this failure is generated by a lack of commitment in property rights. Consider an economy with infinite firms (A, B, C, and so on) and a user i where firm A begins with property rights over user i's data. Firm A uses i's data to produce a superior product, and so is only willing to sell i's data to other firms (or back to i) at a strictly positive price. But, no other firm is willing to pay a positive price for i's data to A without commitment that A won't also resell to another counter-party. Because data is non-rival, no such commitment can be extracted. As such, no transactions can take place over data. This is data hoarding.

Fourth is a concern of servicing **diverse preferences among users**. This is a failure of market thickness.[8] When users have heterogeneous preferences – particularly regarding privacy, ethics, or cultural norms – market-based solutions tend to systematically underserve those with non-mainstream preferences [Waldfogel, 2003]. This is indignifying. HCP addresses this failure by empowering all individuals to define and enforce their own specific preference boundaries through granular controls, ensuring their values are respected regardless of the prevalence of bespoke market solutions.

4.3 Addressing proxy misalignment

The theory of revealed preference states that our preferences are essentially defined by the choices we make [Samuelson, 1938]. This enables inference over actions; an enormously powerful paradigm in a modern world where user action data is bountiful. Yet, modern behavioral work (and common sense) yield many examples where this paradigm may fail – for example, problems of mental accounting (i.e., [Thaler, 1985]) or self control (i.e., [Thaler and Shefrin, 1981]), with both problems exacerbated by the difficulty of inference for complex objectives (a problem of statistics).

In the context of user personalization for general (potentially agentic) tools, this problem is fundamental: current personalization systems rely heavily on behavioral proxies – clicks, time spent, purchase history – rather than direct expressions of user intent.[9] This can create misalignment between what systems optimize for (proxies for user preferences) and what users actually want (true preferences). For instance, a news recommendation system might interpret a user clicking on sensationalist headlines as a preference for such content if they fail to accurately model the user's true mental state. Kleinberg et al. [2024a] identify this as the "inversion problem," where systems must do more than naive inference to produce positive outcomes – they must work backwards from observable actions to infer mental states.

HCP addresses this larger ecosystem challenge by providing a mechanism for **direct preference articulation** to supplement and ground inference. For instance, I may explicitly include a preference against tabloid gossip, thereby using HCP as a commitment mechanism to avoid such news articles. This is particularly valuable for complex, multifaceted preferences that are difficult to infer from behavior alone – such as privacy boundaries, ethical values, or content standards.

4.4 Alignment and value diversity

A core promise of personalized AI is achieving *pluralistic alignment* – systems that are consistent with user values for a wide spectrum of values [Sorensen et al., 2024]. However, current alignment paradigms face significant hurdles. Dominant techniques often rely on feedback from limited, often non-representative rater pools, leading to biased model behavior and narrow value representation [Kirk et al., 2024, Santurkar et al., 2023, Fulay et al., 2024]. Alignment processes like RLHF can inadvertently reduce output diversity and distributional pluralism, risking homogenized responses and an AI monoculture [Durmus et al., 2023].

HCP offers an architectural solution by empowering users with direct control over their preference data. This user-centric model can counteract homogenization by furnishing models with explicit, diverse preference signals. In contrast to existing post-training regimes, HCP supports *steerable pluralism*, where models can be faithfully guided to reflect specific user-defined viewpoints within explicitly identified bounds [Sorensen et al., 2024]. This shifts the alignment paradigm from striving for a single "correct" model behavior to enabling models that can respectfully and capably represent a multitude of user-authorized perspectives, better serving a plural society.

Of course, personalization at the individual level also raises data sufficiency concerns: fine-tuning on a single user's preferences may be noisy or brittle. Conversely, aligning to crowdworker data

[8] Even with an economy of infinitely numbered firms, no firm opens a factory for a single buyer if fixed costs are sufficiently high. The existence of multiple identical buyers would allow firms to defray their fixed costs. This is economies of scale. The problem of niche buyers is that no firm finds sufficient aggregate WTP to cover their costs.

[9] Although AI firms rarely disclose the details of their LLM-tuning pipelines, public product documents already show that they exploit rich behavioral traces. Google's March 2025 announcement of "Gemini with personalization" states that the assistant "will be able to use your Google apps, *starting with your Search history*, to deliver contextually relevant responses" Citron [2025]. Similarly, Meta's January 2025 update notes that "Meta AI can now use your Facebook and Instagram data to personalize its responses" Wiggers [2025].

provides volume but not specificity. A hybrid approach – training on data from demographically or behaviorally "similar" users – may strike a middle ground. HCP enables this possibility by making similarity matching both explicit and user-consented.

5 Limitations and discussion

While HCP offers a promising framework for addressing current challenges in AI personalization, several important considerations must be addressed for successful implementation. In this section, we discuss practical and social challenges in the adoption of HCP. Further ethical considerations are discussed in Section 5.4.

5.1 Practical challenges

The core obstacle is standards convergence. Multiple vendors must agree on a stable interface for declaring, storing, and exchanging preference vectors, yet the pace of model innovation makes any rigid specification brittle. Successful precedents – from TCP/IP to OAuth, HTML – show that interoperability wins when standards are open, modular, and versioned, letting new capabilities slot in without breaking legacy clients [Clark, 1988, Simcoe, 2012, Hardt, 2012, Ghazawneh and Henfridsson, 2013].

An additional, related concern here is bootstrapping adoption incentives. Even *if* designers determine the 'optimal' standards model, one must still convince existing vendors and technologists to embrace them. Incumbents treating preference data as a competitive moat are unlikely to adopt HCP without compelling incentives. Emphasizing HCP's long-term market benefits, recruiting keystone early adopters, and deploying policy nudges can help align incentives before proprietary silos harden.

Yet, for these concerns brought on by a desire to discipline the market, the market may yet be the solution. In particular, if firms can enter which provide preference management solutions superior to those provided by incumbents, then competition between these firms will provide users with the plethora of HCP-solutions desired. Such solutions have the benefit of finding product-market fit in a scoped manner. For example, consider a rollout strategy starting in domains where user context is key, e.g. scheduling. A scheduling HCP could accumulate user context over time (availability rules, constraints, relationships), then expand to adjacent tasks (trip planning, reminders) for more user context, eventually evolving into a general-purpose preference management tool. This approach constitutes one realistic, partial market rollout story by which competition solves the standards problem.

5.2 Risks from deep personalization

While personalization represents one of the most exciting frontiers in AI development, it is crucial to acknowledge potential risks. The very capability that makes personalization valuable – enabling AI systems to adapt profoundly to individual preferences – gives these systems increased purchase on users' lives and decisions. This may magnify risks from bad actors, where models can use increased vectors for belief persuasion towards socially undesirable ends.

Beyond malicious use, user inconsistency also creates direct concerns that require careful oversight. First are off-target effects from **information asymmetry**: users may overlook how a system actually affects their psychology, with recent evidence from sycophancy [Sharma et al., 2025, Fanous et al., 2025]. Second, are concerns from **present bias**. Users may use AI products myopically, becoming attached or dependent to these tools at the detriment of their future well-being.

5.3 Enforceability

Beyond agency concerns, there are also limitations within the larger ecosystem worth considering. While the existence of an HCP following our proposed design could enable an AI user to express their desires for AI model behavior, it does not obligate any AI model to comply, nor does it disable the AI provider's ability to broadly harvest or license user data for its own purposes (including from the HCP). Such a system would need stronger protections – such as those approaching full structured transparency [Trask et al., 2020] – so that users might enforce how their information is

used. Nevertheless, the proof of concept above represents a crucial, informative step towards such a fully enforceable system.

5.4 Ethics and Oversight

Finally, there are also some ethical considerations to note in the (long-term) deployment of HCP.

- **Digital-divide mitigation.** If HCP is usable only by technically sophisticated or affluent users, it risks widening existing inequities in realizing the benefits of technology.
- **Accountability frameworks.** A user-centric architecture needs transparency requirements, audit mechanisms, and accessible dispute-resolution processes to address violations.
- **Social nature of data.** Preferences often have shared or networked ownership; HCP should include governance mechanisms that respect overlapping claims on preference subsets.

These challenges, while significant, are not insurmountable. HCP represents a promising direction for addressing fundamental issues in AI personalization. We view each of the difficulties listed in this section not as insurmountable obstacles, but as research questions worthy of collaborative effort.

6 Future Possibilities

While the immediate benefits of an HCP are substantial, its true potential emerges when we consider the *new* possibilities it enables. This section explores three dimensions of future possibilities: enhanced individual agency, novel downstream mechanisms built upon the preference architecture, and the broader societal implications for markets, policy, and research. For each section below, we offer several, concrete illustrations.

6.1 Expanding User Agency: From Preference Expression to Discovery

Many industries across the information economy are predicated on the importance of preference formation and learning. While academic models often assume stable, well-defined preferences, substantial empirical evidence documents that consumers engage in costly search and experimentation to acquire information about their own utility functions [DellaVigna, 2009]. Classic research on experience goods demonstrates that consumers cannot assess preferences for many products without direct trial [Nelson, 1970], while work on constructive choice processes shows that individuals often form preferences during decision-making rather than retrieving pre-existing ones [Bettman et al., 1998].

HCP enables **systematic preference discovery** through controlled self-experimentation across AI systems and contexts. The HCP becomes not just a repository but a laboratory for exploring one's values and preferences. Such self-experimentation has precedent in the Quantified Self movement (initiated also by a flourishing of personal data) [Swan, 2012], but HCP extends this approach beyond physiological metrics to the domain of values, interests, and decision-making principles. Users can modify their stated preferences, observe downstream AI agent responses across different contexts, and iteratively refine their understanding of their own values and priorities.

We illustrate this potential across several sectors where systematic preference discovery could meaningfully improve user outcomes:

News consumption. Users can experiment systematically with information diet preferences – testing depth versus breadth, source diversity, topic coverage, and analytical framing. Unlike current recommendation systems optimized for engagement, this approach creates space for critical reflection on what news consumption patterns actually serve users' informational goals. Consider a user who is unsure about their preference for news content. They might create different preference profiles – one emphasizing depth and completeness, another brevity and efficiency – and compare their satisfaction with the resulting AI behaviors.

Political information. Testing different information sources and frames to understand political preferences has been central to research on democratic opinion formation [Druckman and Lupia, 2000]. HCP could facilitate deliberate exposure to cross-cutting political information as a method for

reducing polarization, allowing users to explore diverse perspectives while maintaining control over pace and scope:

1. A naive (but useful and immediately implementable) example may be cross-partisan perspective testing. In much the same way individuals can learn about their news preferences, they can explore different ideological lenses – the "conservative fiscal perspective," "progressive Chomsky-ite perspective," or the "libertarian, Hayek-ian focus."
2. A more sophisticated application exploits the reality that political labels (even fine-grained ones!) obscure tremendous heterogeneity in underlying values and priorities. Users can create preference profiles that disaggregate these categories – testing whether they prioritize efficiency over equity, individual autonomy over collective responsibility, or institutional stability over transformative change – across different issues.
3. Finally, one of the most interesting things about political preferences is that, while they may evolve, they are indexed to the same person over *time*. So, an individual may implement experiments on political mores and content specifically against past preferences saved in the HCP. This may take directives of the form, "Set my political preferences on this *new* technology artifact according to the politics I held when I was 14, 24, and 44."

Matching markets. There exist numerous matching markets that clear along individuals preferences and type, where the addition of technology clarifying users preferences would be particularly useful.

Two in particular may be dating [Rosenfeld and Thomas, 2012, Finkel et al., 2012, Hitsch et al., 2010] and school placement (i.e., college admissions or the NRMP match program for residents) [Roth and Peranson, 1999, Gale and Shapley, 1962]. These markets often clear without explicit monetary contracts, making the importance of preference learning and information acquisition particularly important.

Music discovery. Users can systematically explore genre preferences, mood-based listening patterns, and discovery versus familiarity trade-offs across contexts, developing richer self-knowledge beyond the algorithmic recommendations of individual platforms. This mirrors existing efforts in the industry – for example, Spotify's 'New Genres You May Like' initiative, meant to decrease search costs while improving match quality.

Product discovery. Most obviously, the global digital advertising industry – worth over $600 billion annually [eMarketer/Insider Intelligence, 2024] – fundamentally operates on the logic of reducing search frictions and facilitating product discovery. The industry assumes consumers have latent preferences for undiscovered products, with targeted exposure designed to reveal and activate these preferences. HCP could enable an active discovery process from the *user* side, giving users increased latitude in the types of ads they see.

In general, we believe that HCP's greatest utility lies in preference discovery along *high-dimensional* preferences, in markets with substantial product diversity (where expansive search is costly), and in one-shot, high-stakes scenarios – precisely the domains where personalization in the status quo is difficult.

6.2 Novel Downstream Mechanisms: Building on the Preference Layer

Standardizing preference expression and management creates a foundation upon which entirely new mechanisms can develop, much as standardized protocols enabled the flourishing of internet applications by reducing transaction costs and enabling new goods and services [Shapiro and Varian, 1999]. One recent example is the standardization of financial data access through 'open banking' regulations, like Europe's PSD2 [European Union, 2015]. By mandating that incumbent banks share customer data, this policy spurred new firm entry, enabling a wave of downstream fintechs to build innovative financial services on top of the existing infrastructure [Babina et al., 2025]. HCP represents precisely such a standardization of the preference layer, opening possibilities that extend far beyond individual personalization.

The key insight is that when preferences become structured, portable, and machine-readable, they can serve as inputs to coordination mechanisms that were previously impractical due to high transaction costs. Moreover, the mechanisms themselves can include new types of commitment. This enables everything from sophisticated group decision-making to collective bargaining structures that aggregate

individual preferences into coordinated action. Below, we outline several illustrative applications that demonstrate HCP's potential to enable novel forms of digital cooperation and governance.

Guardian assistant systems. Perhaps the most immediately valuable application is the creation of "guardian assistant" layers – middleware AI systems that sit between user HCPs and other digital services. These guardians, operating with full access to user HCPs, serve a crucial dual function. Primarily, they act as digital advocates to enforce a user's own preferences. However, they also serve as a control layer, implementing policies set by trusted third parties that can supersede a user's immediate intentions, either for the user's own protection or to prevent harm to others.

Such guardians could intercept outbound prompts and inbound content to identify persuasive tactics or deceptive patterns, flag potential manipulation attempts based on known user vulnerabilities, filter content, add overlays that provide relevant context, and negotiate automatically with third-party systems based on user-defined boundaries. This protective function is especially vital for children, where a parent's policies for content filtering can override a child's immediate choices to shield them from harmful material.

Furthermore, while much of this paper situates the HCP as a user-specific technology, it's important to note that the 'guardian' can also be used to reflect more complicated social relations. In particular, consider the relation between an employer and employee. A user's calendar data may reveal private information about their employer. To manage this risk, the user's firm may wish to be guardian to their network of employees, overseeing user-specific data scoping to ensure that sensitive firm-specific information isn't accidentally leaked.

Group coordination mechanisms. When individual preferences are structured and accessible, new possibilities emerge for group decision-making that goes far beyond simple polling or majority voting. Tools built atop HCP could aggregate compatible preferences to facilitate coordination problems ranging from scheduling to collaborative project planning. Unlike traditional voting systems, these mechanisms could perform sophisticated preference matching, identifying complementary patterns and potential compromises that satisfy multiple constraints simultaneously [Tessler et al., 2024, Bakker et al., 2022].

Consider planning a group activity where participants have expressed different primary preferences. The system might recognize that while Alice prefers outdoor activities and Bob prefers cultural events, both share a secondary preference for novel experiences – suggesting an outdoor cultural festival as an optimal compromise. This capability explicitly plays out the analogy of revelation mechanisms from economic theory, but with dramatically reduced transaction costs due to the structured preference data that HCP provides. Additionally, it can also be paired with other discursive methods, to enable clearer debate and value negotiation [Burton et al., 2024].

Negotiation, Collective Action. HCP enables users to pool preferences into cooperative structures that can exercise collective leverage, directly addressing fundamental power imbalances in digital markets where individual users face large platforms. Consider a preference cooperative focused on privacy practices: members contribute their privacy preferences to a shared layer, with an agent that negotiates with services on behalf of the entire group. Services might offer improved terms to access this aggregated market, similar to how buying clubs achieve volume discounts through coordinated purchasing power.

This collaborative approach creates collective mechanisms for users to resist surveillance practices and reclaim agency in digital environments [Zuboff, 2023]. Such preference pooling could extend across domains: negotiating improved service terms or features, coordinating responses to services that violate common preference boundaries, facilitating data unions that derive shared value from combined preference data, and creating preference-based mutual aid networks where compatible preferences enable resource sharing.

Education. Educational institutions could provide HCP infrastructure as complementary to the digital investments (e.g., in laptops or tablets) made in schools to support student development. This creates particularly interesting possibilities for controlled agency development in environments where society appropriately limits full autonomy. School districts could maintain HCP structures that lease permissions to students based on developmental appropriateness, gradually expanding student control over their learning preferences as they mature.

Moreover, different pedagogical philosophies could be enacted through preference architectures as well: Montessori approaches might emphasize student choice discovery and self-directed preference formation, while more structured curricula could guide preference development toward specific learning outcomes. The system enables personalized learning at scale while maintaining institutional oversight – teachers gain insight into individual student learning preferences while students gradually develop agency over their educational experience. This represents a technical infrastructure for implementing diverse educational philosophies in ways that can be systematically compared and evaluated.

Democratic governance. HCP could support democratic governance by enabling citizens to share relevant preference profiles with elected representatives or public institutions. Rather than relying on crude polling or responding only to the loudest voices, representatives could access nuanced preference distributions on specific issues, subject to appropriate privacy protections and explicit user consent.

This approach represents a significant evolution beyond current civic technologies. While platforms like Polis enable more nuanced opinion clustering than simple polls, they still require active participation for each issue [Small et al., 2021]. HCP would enable "passive representation"—where citizens' already-articulated values can inform governance without requiring constant civic engagement. This could support deliberative democracy [Fishkin and Luskin, 2005, Burton et al., 2024] by providing governance structures with access to thoughtful citizen input rather than relying solely on vote aggregation or activist mobilization.

Altogether, these possibilities entail just some of the ways that standardizing the preference layer over models may enable new, useful personalization mechanisms.

6.3 Broader Societal Implications

The widespread adoption of HCP (and novel, downstream mechanisms) would likely trigger significant second-order effects across markets, policy, and governance. These implications extend far beyond individual personalization to reshape how digital ecosystems organize around user agency.

Market evolution and new economic structures. Just as app stores emerged atop standardized mobile operating systems, HCP would likely spawn markets for specialized preference management tools, guardian systems, and preference-discovery services. A natural evolution would be the emergence of a "marketplace of licensed guardians" – specialized AI systems certified to protect user interests in specific domains.

These might include **child safety guardians** that enforce age-appropriate interactions based on parent-defined HCP layers, **financial guardians** that protect against manipulation in high-stakes transactions, **health guardians** that ensure medical AI systems respect patient treatment preferences and risk tolerance, and **professional guardians** that maintain workflow preferences while protecting against distractions.

The key idea is that many user contexts have society-approved mores associated with them, but that users' preferred solutions may differ. Such marketplaces would create powerful incentives for innovation in preference protection and enhancement.

Policy development and regulatory frameworks. HCP represents a practical demonstration of data portability and user control that could inform future policy development across multiple domains. By showing that meaningful user control is technically feasible, HCP provides policymakers with a concrete reference model for regulations concerning data rights and AI governance. Current regulatory frameworks like GDPR include data portability requirements, but these remain largely theoretical without practical implementations. HCP offers a template for "by-design" approaches to regulation [Mulligan et al., 2016] – embedding policy objectives directly into technical architecture rather than imposing them through external compliance requirements.

Distributed governance models. The preference layer architecture suggests a novel model for distributed governance of AI systems, where control is exercised not through centralized oversight but through the aggregated preferences of users themselves. This approach aligns with concepts of regulation by architecture, where technical design choices enforce normative objectives [Lessig, 2009]. Rather than relying solely on top-down regulatory intervention, HCP enables bottom-up governance

through collective user agency – a form of technological democracy where the architecture itself becomes a mechanism for expressing and enforcing societal values about AI behavior.

6.4 Legal and Regulatory Imperatives

Beyond arguments from market incentives or user agency, a strong case can be made that an architecture like HCP is also an emerging legal imperative under major data protection regimes. Both Europe's General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA) establish a robust right to data portability. Specifically, Art. 20 §§1–2 of the GDPR grants individuals the right to receive their personal data in a "structured, commonly used and machine-readable format" and to transmit it to another service provider without hindrance [European Parliament and Council of the EU, 2016]. The CCPA mirrors this (in Cal. Civ. Code § 1798.100(d)), requiring businesses to provide personal information in a portable, readily usable format [California Legislature, 2018].

The data used for AI personalization, whether explicitly stated preferences or behavioral patterns, is likely covered by these regulations. This data falls under the GDPR's definition of personal data so long as it is linked to an identifiable person, and is similarly protected under the CCPA if it "identifies, relates to, describes, is capable of being associated with, or could reasonably be linked" to a specific individual. Consequently, the siloed, provider-centric models that lock in this information and create high switching costs violate the requirements of these privacy laws. An interoperable, user-controlled system like HCP is therefore not just a foundation for a more competitive market but a necessary technical prerequisite for AI providers to fulfill their legal obligations, ensuring users can meaningfully exercise their right to port their digital identity and preferences across services.

7 Conclusion

As generative AI technologies become more capable and widespread, the mechanisms for personalization become increasingly consequential, shaping not just user experience but also the ability to coordinate and communicate at scale. In this paper, we have argued that current approaches – where preferences are inferred rather than expressed, controlled by providers rather than users, and fragmented across services rather than portable – fail to realize the full potential of personalization while introducing significant risks of manipulation, privacy violation, and lock-in.

The central thesis of this position paper is that the architectural locus of control over preference data matters profoundly for personalization, and we have delivered design principles to guide new solutions for pluralistic alignment. HCP offers a path forward: an architecture that enables seamless portability across services, supports rich articulation of complex preference structures, and prevents lock-in without sacrificing personalization. This is not merely a technical proposal but a reimagining of the relationship between users and AI systems, grounded in principles of autonomy, transparency, and productive competition.

Beyond these immediate benefits, HCP opens transformative possibilities across three interconnected dimensions.

1. First, enhanced individual agency transforms preference management from passive expression to active discovery, where users experiment with different preference profiles to better understand their values through iterative refinement. This, in particular, may also enable more complex "guardian assistant" AI layers, where socio-political constraints interact with this preference discovery.
2. Second, novel collective mechanisms emerge from standardized preference expression. These include sophisticated group decision-making tools that identify complementary preferences and optimal compromises, building on social choice theory made practically implementable at scale. Users can form preference cooperatives "to negotiate collectively with services, addressing power imbalances and creating what scholars term a right to sanctuary" in digital environments [Zuboff, 2023]. HCP could even support new forms of democratic governance where citizens share nuanced preference profiles with public institutions, advancing deliberative democracy [Burton et al., 2024].
3. Third, broader ecosystem implications, particularly along new markets and policy possibilities. For each of the new mechanisms, there exist new market possibilities for solutions

to compete in and provide improved user personalization. The increased scope of this personalization – and the increased purchase this personalization buys on our actions – intensifies the policy imperative for safe and aligned AI.

The path forward requires coordinated effort across technical development, policy innovation, and social adoption, but the potential rewards – a pluralistic, user-empowered AI ecosystem – justify the coordination challenges ahead. These possibilities position HCP as a foundation for reimagining not just individual AI interactions but collective digital life and governance – fostering an ecosystem that genuinely reflects and respects the diversity of human values.

Acknowledgments

We are grateful to Alan Chan, Tantum Collins, Dazza Greenwood, Thomas Hardjono, Galen Hines-Pierce, John Horton, Joshua Levy, and Robert Mahari for helpful conversations and feedback on earlier drafts.

References

Alessandro Acquisti and Hal R Varian. Conditioning prices on purchase history. *Marketing Science*, 24(3):367–381, 2005.

Christopher Allen. The path to self-sovereign identity. `https://www.lifewithalacrity.com/2016/04/the-path-to-self-sovereereign-identity.html`, 2016.

Anthropic. Introducing the model context protocol, 11 2024. URL `https://www.anthropic.com/news/model-context-protocol`. Accessed: 2025-05-18.

Tania Babina, Saleem Bahaj, Greg Buchak, Filippo De Marco, Angus Foulis, Will Gornall, Francesco Mazzola, and Tong Yu. Customer data access and fintech entry: Early evidence from open banking. *Journal of Financial Economics*, 169:103950, 2025.

Michiel Bakker, Martin Chadwick, Hannah Sheahan, Michael Tessler, Lucy Campbell-Gillingham, Jan Balaguer, Nat McAleese, Amelia Glaese, John Aslanides, Matt Botvinick, et al. Fine-tuning language models to find agreement among humans with diverse preferences. *Advances in Neural Information Processing Systems*, 35:38176–38189, 2022.

James R. Bettman, Mary Frances Luce, and John W. Payne. Constructive consumer choice processes. *Journal of Consumer Research*, 25(3):187–217, 1998. doi: 10.1086/209535.

Jason W Burton, Ezequiel Lopez-Lopez, Shahar Hechtlinger, Zoe Rahwan, Samuel Aeschbach, Michiel A Bakker, Joshua A Becker, Aleks Berditchevskaia, Julian Berger, Levin Brinkmann, et al. How large language models can reshape collective intelligence. *Nature human behaviour*, 8(9): 1643–1655, 2024.

California Legislature. California civil code §1798.100(d). California Consumer Privacy Act, 2018. URL `https://leginfo.legislature.ca.gov`. Subdivision(d): business obligations on consumer requests.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: Building production-ready ai agents with scalable long-term memory. *arXiv preprint arXiv:2504.19413*, 2025.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems 30*, pages 4299–4311, 2017.

Dave Citron. Gemini gets personal, with tailored help from your google apps, 2025. URL `https://blog.google/products/gemini/gemini-personalization/`. Accessed 2025-05-21.

David D. Clark. The design philosophy of the darpa internet protocols. *ACM SIGCOMM Computer Communication Review*, 18(4):106–114, 1988. doi: 10.1145/52325.52336.

Ronald H. Coase. The problem of social cost. *Journal of Law and Economics*, 3(1):1–44, 1960. doi: 10.1086/466560. URL `https://www.journals.uchicago.edu/doi/10.1086/466560`.

Stefano DellaVigna. Psychology and economics: Evidence from the field. *Journal of Economic Literature*, 47(2):315–372, 2009. doi: 10.1257/jel.47.2.315.

James N. Druckman and Arthur Lupia. Preference formation. *Annual Review of Political Science*, 3: 1–24, 2000. doi: 10.1146/annurev.polisci.3.1.1.

Esin Durmus, Karina Nguyen, Thomas I Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, et al. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*, 2023.

eMarketer/Insider Intelligence. Global and us digital ad spending forecast 2024, 2024. URL https://www.emarketer.com/content/digital-ad-spend-worldwide-pass-600-billion-this-year. Accessed 27 May 2025.

European Parliament and Council of the EU. Regulation (eu) 2016/679 (general data protection regulation), 2016. URL https://eur-lex.europa.eu/eli/reg/2016/679/oj. Art. 20: Right to data portability.

European Union. Directive (EU) 2015/2366 of the European Parliament and of the Council of 25 November 2015 on payment services in the internal market. Official Journal of the European Union, L 337/35, dec 2015. URL http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:32015L2366. Commonly known as the second Payment Services Directive (PSD2).

Aaron Fanous, Jacob Goldberg, Ank A. Agarwal, Joanna Lin, Anson Zhou, Roxana Daneshjou, and Sanmi Koyejo. Syceval: Evaluating llm sycophancy, 2025. URL https://arxiv.org/abs/2502.08177.

Joseph Farrell and Paul Klemperer. Coordination and lock-in: Competition with switching costs and network effects. *Handbook of industrial organization*, 3:1967–2072, 2007.

Eli J Finkel, Paul W Eastwick, Benjamin R Karney, Harry T Reis, and Susan Sprecher. Online dating: A critical analysis from the perspective of psychological science. *Psychological Science in the Public interest*, 13(1):3–66, 2012.

James S Fishkin and Robert C Luskin. Experimenting with a democratic ideal: Deliberative polling and public opinion. *Acta politica*, 40:284–298, 2005.

Suyash Fulay, William Brannon, Shrestha Mohanty, Cassandra Overney, Elinor Poole-Dayan, Deb Roy, and Jad Kabbara. On the relationship between truth and political bias in language models. *arXiv preprint arXiv:2409.05283*, 2024.

David Gale and Lloyd S Shapley. College admissions and the stability of marriage. *The American mathematical monthly*, 69(1):9–15, 1962.

Ahmad Ghazawneh and Ola Henfridsson. Balancing platform control and external contribution in third-party development: The boundary resources model. *Information Systems Journal*, 23(2): 173–192, 2013. doi: 10.1111/j.1365-2575.2012.00406.x.

Google. Gemini apps privacy hub. https://support.google.com/gemini/answer/13594961, 2025. Last updated May 20, 2025.

Google. Gemini with ai personalization — get help made just for you, 2025. URL https://gemini.google/overview/personalization/. Accessed 2025-05-21.

Google. Agent2agent (a2a) protocol, 2025. URL https://github.com/google/A2A.

Sanford J Grossman and Oliver D Hart. The costs and benefits of ownership: A theory of vertical and lateral integration. *Journal of Political Economy*, 94(4):691–719, 1986.

John Hagel III and Jeffrey F Rayport. The coming battle for customer information. *The McKinsey Quarterly*, page 64, 1997.

Dick Hardt. The oauth 2.0 authorization framework, October 2012. URL https://www.rfc-editor.org/info/rfc6749.

Gunter J Hitsch, Ali Hortacsu, and Dan Ariely. Matching and sorting in online dating. *American Economic Review*, 100(1):130–163, 2010.

Charles I Jones and Christopher Tonetti. Nonrivalry and the economics of data. *American Economic Review*, 110(9):2819–2858, 2020.

M. Jones, J. Bradley, M. Machulak, and P. Hunt. OAuth 2.0 Dynamic Client Registration Protocol, July 2015. URL https://datatracker.ietf.org/doc/html/rfc7591. IETF Standard RFC7591.

Hannah Rose Kirk, Bertie Vidgen, Paul Röttger, and Scott A Hale. The benefits, risks and bounds of personalizing the alignment of large language models to individuals. *Nature Machine Intelligence*, 6(4):383–392, 2024.

Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Manish Raghavan. The inversion problem: Why algorithms should infer mental state and not just predict behavior. *Perspectives on Psychological Science*, 19(5):827–838, 2024a.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. The challenge of understanding what users want: Inconsistent preferences and engagement optimization. *Management science*, 70(9): 6336–6355, 2024b.

Kenneth C Laudon. Markets and privacy. *Communications of the ACM*, 39(9):92–104, 1996.

Lawrence Lessig. *Code: And other laws of cyberspace*. ReadHowYouWant. com, 2009.

Xinyu Li, Ruiyang Zhou, Zachary C. Lipton, and Liu Leqi. Personalized language modeling from personalized human feedback, 2024. URL https://arxiv.org/abs/2402.05133.

Meta. Meta privacy policy. https://www.facebook.com/privacy/policy/, 2025. Accessed May 2025.

Meta. Building toward a smarter, more personalized assistant, 2025. URL https://about.fb.com/news/2025/01/building-toward-a-smarter-more-personalized-assistant/. Accessed 2025-05-21.

Deirdre K Mulligan, Colin Koopman, and Nick Doty. Privacy is an essentially contested concept: a multi-dimensional analytic for mapping privacy. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2083):20160118, 2016.

Alexander Mühle, Andreas Grüner, Tatiana Gayvoronskaya, and Christoph Meinel. A survey on essential components of a self-sovereign identity. *Computer Science Review*, 30:9–29, 2018.

Philip Nelson. Information and consumer behavior. *Journal of Political Economy*, 78(2):311–329, 1970. doi: 10.1086/259630.

OpenAI. Privacy policy (rest of world). https://openai.com/policies/row-privacy-policy/, 2024. Updated November 4, 2024.

OpenAI. The power of personalized ai, 2025. URL https://openai.com/global-affairs/the-power-of-personalized-ai/. Accessed 2025-05-21.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*, 36(7):3580–3599, 2024.

Sriyash Poddar, Yanming Wan, Hamish Ivison, Abhishek Gupta, and Natasha Jaques. Personalizing reinforcement learning from human feedback with variational preference learning, 2024. URL https://arxiv.org/abs/2408.10075.

Antti Poikola, Kai Kuikkaniemi, and Harri Honko. *MyData – A Nordic Model for human-centered personal data management and processing*. Ministry of Transport and Communications Finland, 2015. URL https://julkaisut.valtioneuvosto.fi/handle/10024/78439.

Valentina Pyatkin, Jena D Hwang, Vivek Srikumar, Ximing Lu, Liwei Jiang, Yejin Choi, and Chandra Bhagavatula. Clarifydelphi: Reinforced clarification questions with defeasibility rewards for social and moral situations. *arXiv preprint arXiv:2212.10409*, 2022.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023.

Michael J Rosenfeld and Reuben J Thomas. Searching for a mate: The rise of the internet as a social intermediary. *American Sociological Review*, 77(4):523–547, 2012.

Alvin E Roth and Elliott Peranson. The redesign of the matching market for american physicians: Some engineering aspects of economic design. *American economic review*, 89(4):748–780, 1999.

Andrei Vlad Sambra, Essam Mansour, Sandro Hawke, Maged Zereba, Nicola Greco, Abdurrahman Ghanem, Dmitri Zagidulin, Ashraf Aboulnaga, and Tim Berners-Lee. Solid: a platform for decentralized social applications based on linked data. *MIT CSAIL & Qatar Computing Research Institute, Tech. Rep.*, 2016, 2016.

Paul A. Samuelson. A Note on the Pure Theory of Consumer's Behaviour. *Economica*, 5(17):61–71, February 1938. doi: 10.2307/2548836.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? In *International Conference on Machine Learning*, pages 29971–30004. PMLR, 2023.

Carl Shapiro and Hal R Varian. The art of standards wars. *California management review*, 41(2): 8–32, 1999.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models, 2025. URL `https://arxiv.org/abs/2310.13548`.

Timothy S. Simcoe. Standard setting committees: Consensus governance for shared technology platforms. *American Economic Review*, 102(1):305–336, 2012. doi: 10.1257/aer.102.1.305.

Christopher Small, Michael Bjorkegren, Timo Erkkilä, Lynette Shaw, and Colin Megill. Polis: Scaling deliberation by mapping high dimensional opinion spaces. *Recerca: revista de pensament i anàlisi*, 26(2), 2021.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, et al. A roadmap to pluralistic alignment. *arXiv preprint arXiv:2402.05070*, 2024.

Tobin South, Samuele Marro, Thomas Hardjono, Robert Mahari, Cedric Deslandes Whitney, Dazza Greenwood, Alan Chan, and Alex Pentland. Authenticated delegation and authorized ai agents. *Forty-Second International Conference on Machine Learning*, 2025.

Melanie Swan. Sensor mania! the internet of things, wearable computing, objective metrics, and the quantified self 2.0. *Journal of Sensor and Actuator networks*, 1(3):217–253, 2012.

Michael Henry Tessler, Michiel A Bakker, Daniel Jarrett, Hannah Sheahan, Martin J Chadwick, Raphael Koster, Georgina Evans, Lucy Campbell-Gillingham, Tantum Collins, David C Parkes, et al. Ai can help humans find common ground in democratic deliberation. *Science*, 386(6719): eadq2852, 2024.

Richard Thaler. Mental accounting and consumer choice. *Marketing science*, 4(3):199–214, 1985.

Richard H Thaler and Hersh M Shefrin. An economic theory of self-control. *Journal of political Economy*, 89(2):392–406, 1981.

Andrew Trask, Emma Bluemke, Teddy Collins, Ben Garfinkel, Eric Drexler, Claudia Ghezzou Cuervas-Mons, Iason Gabriel, Allan Dafoe, and William Isaac. Structured transparency: a framework for addressing use/mis-use trade-offs when sharing information. *CoRR*, abs/2012.08347, 2020. URL `https://arxiv.org/abs/2012.08347`.

Hal R. Varian. Economic aspects of personal privacy. In *Privacy and Self-Regulation in the Information Age*. U.S. Department of Commerce, 1996.

Hal R Varian, Joseph Farrell, and Carl Shapiro. *The economics of information technology: An introduction*. Cambridge University Press, 2004.

V Brian Viard. Do switching costs make markets more or less competitive? the case of 800-number portability. *The RAND Journal of Economics*, 38(1):146–163, 2007.

Joel Waldfogel. Preference externalities: An empirical study of who benefits whom in differentiated-product markets. *RAND Journal of Economics*, 34(3):557–568, 2003.

Samuel D. Warren and Louis D. Brandeis. The right to privacy. *Harvard Law Review*, 4(5):193–220, 1890. URL `https://www.jstor.org/stable/1321160`. Accessed via JSTOR.

Alan F Westin. Privacy and freedom. *Washington and Lee Law Review*, 25(1):166, 1968.

Kyle Wiggers. Meta ai can now use your facebook and instagram data to personalize its responses, January 2025. URL https://techcrunch.com/2025/01/27/meta-ai-can-now-use-your-facebook-and-instagram-data-to-personalize-its-responses/. TechCrunch, accessed 23 May 2025.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, and Paul Christiano. Fine-tuning language models from human preferences, 2019.

Shoshana Zuboff. The age of surveillance capitalism. In *Social theory re-wired*, pages 203–213. Routledge, 2023.

Guy Zyskind, Oz Nathan, et al. Decentralizing privacy: Using blockchain to protect personal data. In *2015 IEEE security and privacy workshops*, pages 180–184. IEEE, 2015.

Guy Zyskind, Tobin South, and Alex Pentland. Don't forget private retrieval: distributed private similarity search for large language models. In *Proceedings of the Fifth Workshop on Privacy in Natural Language Processing*, pages 7–19, 2024.