# ROBUST AI PERSONALIZATION CONTROLS: THE HUMAN CONTEXT PROTOCOL

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

**Tobin South**
Stanford University
tsouth@stanford.edu

**Talfan Evans**
Cursive
talfan@cursive.ai

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

**Jiaxin Pei**
Stanford University
pedropei@stanford.edu

**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

Personalization underpins the modern digital economy. Today, personalization is largely implemented through provider-managed infrastructure that infers user preferences from behavioral data, with limited portability or user control. However, large language models (LLMs) are increasingly being used to perform tasks on users' behalf. The age of LLMs for the first time provides a path to a more controllable and interpretable personalization paradigm, grounded in user-expressed natural language preferences and context. We propose the Human Context Protocol (HCP), a user-centric approach to representing and sharing personal preferences across AI systems.[1] HCP treats preferences as a portable, user-governed layer in the personalization stack, enabling interoperability, scoped access, and revocation. Along with a working prototype to ground discussion, we consider adoption dynamics and market incentives, high-stakes use cases, and outline novel paths via the HCP towards trustworthy personalization in the human-AI economy.

## 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]. This evolution toward personalization is accelerating rapidly, with major AI providers including OpenAI [2025a], Google [2025], and Meta Meta [2025] having announced personalization features in 2025 as central axes of their development roadmaps.

However, the current paradigm for personalization presents significant challenges. First, preference data is often opaque and incontestable – while personalization is predicated on knowing user preferences, users rarely see what the system "knows," exacerbating problems of privacy, and of shallow or inaccurate personalization [Kleinberg et al., 2024]. Second, preference data is often non-portable, reinforcing user lock-in and harming market competition. Context cannot move easily across models or services, which raises switching costs and stymies downstream interoperability [Farrell and Klemperer, 2007]. Both challenges reflect deeper questions about user ownership and portability of preference data, and about the thin, provider-dominated market for personalization infrastructure.

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 standardizing access to context, MCP does not address

[1] Project site: www.hcp.me.

questions of ownership, granular user control, or privacy management for personal preferences. Yet these questions are vital to deployment.

We argue that a dedicated, user-centric layer for preference management is a core requirement for building AI systems that are genuinely personal, interoperable, and aligned with diverse human values. To this end, we propose the Human Context Protocol (HCP), a system in which user preferences are managed by a dedicated intermediary – e.g., an LLM – that serves as the interface between individuals and the AI systems acting on their behalf.

Concretely, an HCP should enable individuals to:

- **Control access** to their preferences across LLM-powered services through fine-grained, revocable, and purpose-scoped permissions;
- **Port preferences** across models and providers, reducing switching costs and mitigating lock-in; and
- **Actively shape** how preferences inform model behavior via clear, in-context elicitation and correction loops.

The design of an HCP for user context addresses a market gap: current providers have weak incentives to enable portability or relinquish custody of preference data. A user-centric substrate – separate from any single model provider – establishes the missing infrastructure for private self-management of preferences upon which personalization can be built.

The paper proceeds as follows. In §2 we describe the background and related work. In §3, we propose the HCP. In §4 we consider arguments for the HCP and potential limitations. §5 considers particular human-AI use cases and §6 concludes.

## 2 Background and related work

The conversation on digital personalization often begins with the countervailing right to privacy. This 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 [Westin, 1968]. Yet, as online data proliferated in the computer age – 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 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 this history 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 [Ouyang et al., 2022; Poddar et al., 2024; Sorensen et al., 2025].

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.

[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.

Table 1: Evolution of User Data Control.

| Initiative | Key Idea | Mechanism | Limitations |
|---|---|---|---|
| Infomediaries<br>(Late 1990s) | Brokered user data via intermediaries | Third-party agents managing consent | Indirect control; user frictions; requires large market adoption |
| MyData<br>(2010s) | Data sovereignty as a civic right | Normative principles | Lacked a specific technical implementation |
| Solid Project<br>(Mid 2010s) | User-controlled decentralized storage | Data "pods" with revocable permissions | User frictions (self-hosting); ecosystem still developing; limited natural language scoping |
| SSI<br>(Mid 2010s) | Portable, user-owned digital identity | DIDs and verifiable credentials | Limited to identity attestations; architecturally unsuited for rich data |
| Web3 Data Wallets<br>(Late 2010s) | Custodial control over digital assets | Keys, smart contracts, blockchain | High user frictions; limited legal recourse (relies on "code is law"); asset-centric design |
| **HCP**<br>**(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

Personalization in contemporary AI systems does not arise from a user-mediated preference layer. Instead, it is implemented through mechanisms embedded within model-provider infrastructure. While these mechanisms differ in how preferences are obtained, they share two structural features: (i) preference representations remain internal to providers, and (ii) personalization does not carry across services. In this section, we outline how personalization in AI is currently implemented and why these methods, taken together, are structurally limited.

**Post-training.**
All commercially deployed models undergo post-training, typically combining supervised fine-tuning with population-level preference-based methods such as reinforcement learning from human feedback (RLHF) or direct preference optimization (DPO) [Ouyang et al., 2022; Rafailov et al., 2023]. These procedures encode broad behavioral priors – such as instruction-following and safety – directly into model parameters.

Providers also introduce explicit values and guardrails at this stage. Anthropic's "helpful, honest, harmless" (HHH) framework and Constitutional AI are canonical examples [Askell et al., 2021; Bai et al., 2022], as well as OpenAI's Model Spec [OpenAI, 2025b]. These choices produce a pliant model for downstream users and a common baseline across users.

**In-context personalization, user memory.**
Most user-visible personalization occurs through in-context specification. By stating preferences or situational context in natural language (e.g., "I like Korean food"), users can elicit tailored responses. This form of personalization is legible and flexible, but confined to the active context window: once a session resets, the information is lost unless retained elsewhere.

To persist context across interactions, providers deploy memory systems that store fragments of prior conversations, often supplemented by persistent fields such as OpenAI's "Custom Instructions."[3] Retrieved memories and retrieval-augmented generation (RAG) further personalize responses by supplying external context – documents, calendars, or emails – via tool calls. These mechanisms improve continuity, but stored preferences remain unstructured, provider-managed, and unavailable outside the originating system.

[3] https://openai.com/index/custom-instructions-for-chatgpt/.

**Behavioral inference.**
A distinct approach infers preferences from behavior rather than eliciting them directly. Systems construct implicit representations from interaction patterns, choices, or engagement signals. Recent work formalizes this using latent-variable models, where inferred user codes condition reward models and downstream policies [Poddar et al., 2024; Li et al., 2024]. Earlier dialog systems employed similar techniques, deriving multiple reward functions from offline conversational logs and optimizing policies against those inferred signals [Jaques et al., 2019, 2020]. These inferred representations remain internal to provider optimization pipelines. Users do not directly observe, edit, or transfer them, and their influence is mediated by provider-defined objectives.

**Explicit elicitation.**
Some systems attempt explicit preference elicitation through surveys, critique loops, or reflective dialogue [Blair et al., 2025; Handa et al., 2025]. These approaches can surface preferences that are difficult to infer passively. However, elicited information is incorporated into the same provider-controlled training or inference pipelines, rather than exposed as a persistent object under user control.

**Structural pattern.**
Across these approaches the same pattern recurs. Preferences are formed and acted upon within provider systems, but do not exist as durable, user-governed representations.

This architecture has economic consequences. Although users typically retain nominal rights to delete their data or opt out of certain uses, model providers often maintain broad, perpetual license to use, modify, and sublicense user-provided information, including preference-relevant signals. In practice, this grants providers substantial de facto control over preference data, in the sense of residual rights emphasized by Grossman and Hart [1986]. Users may supply the inputs, but they do not control how those inputs are combined, interpreted, or deployed across contexts and services.

Note that nothing in this diagnosis requires a single representational format. Preference should exist as a distinct layer in the personalization stack – one that competing implementations can plug into. Natural language is the most obvious candidate: it is legible to users, interoperable, and native to language models. But it need not be the only modality. What matters is that preference representations, whatever their form, are user-mediated, portable, and subject to explicit authorization if users desire. The next section discusses these desiderata for system design.

## 3 Human Context Protocol (HCP)

The goal of the HCP is to be an easy and reliable control layer that governs how AI systems access personal user context. Its defining feature is a dedicated intermediary – most naturally, the *HCP LLM* – which interprets user preferences and enforces scoped, minimal disclosure to downstream models at inference time.

### 3.1 Key attributes

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

- **Interoperability**: 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.

- **Encapsulation**: For HCP to provide genuine utility, it must be capable of richly capturing user preferences. While current solutions have limitations in preference elicitation and representation, the HCP's data model should leverage advances in preference representation (whether as text, graph-based knowledge structures, vector embeddings), provided those representations are portable.

- **Control**: 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. This aligns with the principle of data minimization (as in GDPR [2016] Article 5(1)c), ensuring only necessary information is disclosed for a given query.

- **Security**: 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., 2025a].

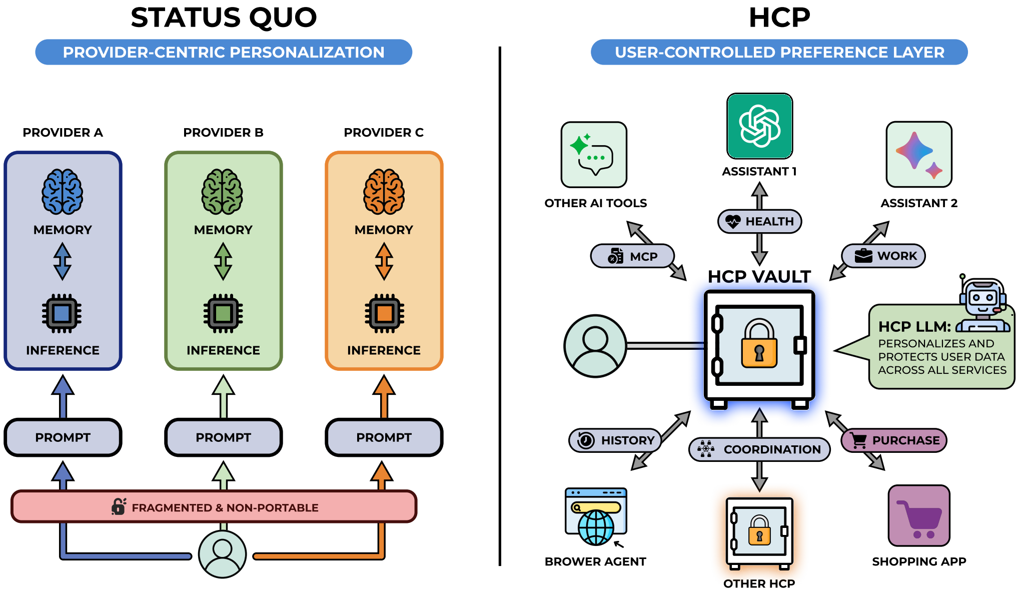

Figure 1: Illustration of two personalization paradigms. On the left panel, personalization is implemented within individual AI services, with user prompts and interaction data feeding service-specific memory and inference pipelines. As a result, preferences are non-portable across providers. On the right panel, a shared preference layer is introduced in which user context is stored in a centralized HCP vault and selectively accessed by multiple assistants, tools, and applications through an HCP-mediated interface. 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.

### 3.2 System design

This paper does not prescribe a definitive implementation for HCP; any system that satisfies the aforementioned design attributes would be suitable. However, to facilitate discussion, we outline a *potential* protocol architecture below. The subsections describe preference representation, supported actions, initialization, and mechanisms for access control, security, and authentication.

#### 3.2.1 Context representation and storage

One key aspect of the HCP system design concerns the representation and storage of personal context and preferences. Language models operate natively in text with high fidelity in both reading and writing. This makes text a well-suited paradigm for storing and editing preferences – unlike, say, large-scale recommender systems, which typically encode user preferences as task-specific embeddings learned implicitly from behavior and tightly coupled to a particular model or objective. This format aligns with the dominant input modalities of most LLMs, offering broad interoperability without requiring specialized serialization formats or custom embeddings. For example, preferences such as "I prefer Mediterranean cuisine" or "I enjoy movies directed by Christopher Nolan" are both human-readable and machine-usable.

To reduce ambiguity, preferences may be enriched with lightweight schema annotations (e.g., a category tag like `food` or a confidence score). Versioning should be supported to allow preferences to evolve over time while preserving historical context. Clarification protocols are also needed to let users refine vague or underspecified statements interactively [Pyatkin et al., 2022].

The underlying storage mechanism could vary: a single comprehensive document (e.g., JSON or Markdown) serving as a unified profile; a key-value store with attached access controls; a graph-based representation capturing relationships between preferences [Pan et al., 2024]; or integration with personal data pods like Solid [Sambra et al., 2016]. For larger datasets, vector databases supporting semantic retrieval can make scalability cheap.

### 3.2.2 Actions

The HCP must support a complete lifecycle of preference operations. At a minimum, four types of actions are envisioned:

- **Read**: retrieving preferences on demand, possibly filtered by category or queried semantically (e.g., "What are the user's vacation preferences?").
- **Write**: adding new preferences, with policies for conflict resolution when overlapping entries exist.
- **Update**: modifying preferences while maintaining version history.
- **Delete**: removing preferences, with compliance guarantees for data erasure such as GDPR's "right to be forgotten."

These actions may be exposed as RESTful endpoints (`GET`, `POST`, `PUT`, `DELETE`) or as MCP tools callable directly by LLM agents. A central design question is granularity. One pragmatic design is to use general-purpose methods like `searchPreferences` and `updatePreferences`, with explicit scoping parameters (e.g., `category=food`) that control retrieval and authorization.

### 3.2.3 Initialization

Initialization determines how an HCP instance is created and begins managing personal context. Bootstrapping could draw on onboarding interviews, user-supplied documents, software integrations, or imported digital traces such as playlists and browsing history. Over time, the preference corpus should evolve through continuous updates, balancing recency with user control. This is similar to how many existing application-specific memory systems work for chatbots.

The HCP LLM plays a central role during initialization and beyond. This model, potentially smaller, locally hosted, or specialized, begins by making common-sense decisions to be later refined or with a minimal user intake process.

### 3.2.4 Access control, security, and authentication

Given the sensitivity of user preference data, an implementation must enforce strong guarantees of confidentiality, integrity, and accountability.

**Access Control** Users require fine-grained, revocable permissions. For example, dietary preferences may be shared with a recipe generator without disclosing unrelated medical data. Control must extend to temporary versus persistent sharing, with the ability to revoke access at any time. Revocation should propagate downstream, potentially requiring compliance actions such as deletion or model retraining rollback.

**Security** All preferences must be encrypted at rest and in transit. Where higher assurance is required, users may hold their own encryption keys, supplying credentials only at runtime. A zero-trust architecture could be used, preventing services from implicitly inheriting unnecessary access. Transparent audit logs allow users to review exactly which agents accessed which data and when.

**Authentication and Integrity** Strong authentication and authorization (e.g., OAuth 2.0) is essential for all external agents, connecting agent or HCP actions to user invocations [South et al., 2025b]. Integrity protections, such as digital signatures, ensure that preference data cannot be tampered with. In addition, HCP must anticipate emerging threats in LLM contexts: inference attacks that reconstruct hidden preferences from observable outputs; adversarial prompting, where malicious models attempt to elicit oversharing; and subtle manipulation of smaller HCP LLMs by more capable external systems.

In sum, access control and security should be designed around principles of *data minimization*, *least privilege*, and *user sovereignty*, positioning HCP as a trustworthy, auditable control layer for managing personal context.

### 3.2.5 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.[4]

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

[4] For a proof-of-concept, see *https://hcp.me/poc-repo*.

authentication and granular authorization for distinct preference categories. For simplicity, this demonstration omits the orchestrating LLM – instead relying upon access requests passed through MCP to determine what information is shared with the user. More complex implementations, which ingest existing personal context and address the cold start problem to route relevant information, are also available.[5]

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 Example

**Concrete user scenario.** *Alex* uses a Claude-based assistant on their phone. They install the HCP integration by adding the HCP–MCP server from the assistant's integrations marketplace and completing an OAuth 2.0 consent flow that grants *category-scoped* access to `outdoor_gear` and `health_context` only. During a chat about hiking, Alex mentions: "I need to make sure my boots support high arches." The assistant recognizes this as a standing constraint and invokes `addPreferences` on the HCP MCP tool, which stores a natural-language entry under `health_context` with provenance (time, source message, model).

Two weeks later, Alex asks: "What lightweight boots should I buy for the Pacific Crest Trail?" The assistant calls `searchPreferences`. The *HCP LLM* evaluates the request against the authorized categories, retrieves only the minimal relevant snippets (e.g., the "high arch support" constraint), and returns a scoped preference bundle. The assistant may then query external product catalogs filtering for arch support and compose a response. Throughout, HCP enforces least-privilege, maintains an audit trail, and allows Alex to revoke scopes or edit entries at any time.

**End-to-end flow.**

1. **Setup.** User enables HCP via MCP integration [Anthropic, 2024]; OAuth 2.0 consent is issued with category-scoped grants [Jones et al., 2015].
2. **Capture.** Assistant tool-call `addPreferences` sends {`category=outdoor_gear`, NL text, metadata} to HCP.
3. **Persist.** HCP validates token and scope, normalizes/versions the NL entry, and stores it in the preference store (with optional vector index).
4. **Query.** Later, assistant invokes `searchPreferences` for "shoes for the Pacific Crest Trail."
5. **Minimize.** HCP validates scope; the HCP LLM selects only relevant items from authorized categories and redacts unrelated fields.
6. **Compose.** Assistant may call external product APIs or A2A frameworks [Google, 2025] using only derived, minimized preference facts.
7. **Explain & Log.** Response includes rationale and (optionally) HCP-provided citations; HCP appends an auditable record. User can view, edit, or revoke.

**Interoperability note.** While the example uses MCP tooling for assistant integration and OAuth 2.0 for authorization, the same flow applies with alternative agent-to-agent transports [Google, 2025] or assistant runtimes. The key invariant is that HCP mediates preference access, applies data minimization at inference time via the HCP LLM, and preserves user control through explicit, revocable scopes and auditable operations.

## 4 Discussion

### 4.1 A Coasian benchmark: does architecture matter?

This paper argues that designers ought to build a preference layer that exists on top of (and independent of) large model providers. Yet, 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]. Hence, 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.

[5] For a more complex implementation, see *github.com/loyalagents/context-router*

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 market failures drive a wedge between theory and reality.

### 4.1.1 Why the benchmark fails for AI personalization

Below, we describe some particular market failures which refute the naive intuition from Coase's theorem.

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]. A second illustrative example comes from financial services – prior to open banking regulations, consumers' transaction histories were locked within incumbent banks, creating high switching costs and limiting entry by new financial intermediaries. Open banking regimes, such as Europe's PSD2, mandated standardized, user-authorized access to account data, enabling third-party providers to compete on services while banks retained custody of funds. Empirically, these reforms increased fintech entry and competition [Babina et al., 2025]. 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.

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.

## 4.2 Addressing proxy misalignment

The theory of revealed preference states that our preferences are essentially defined by the choices we make [Samuelson, 1938]. This justifies 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], [Laibson, 1997]), with both problems exacerbated by the difficulty of inference for complex objectives (a problem of statistics).

[6] See also the discussion at the end of 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.

[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.

In the context of user personalization for general (potentially agentic) tools, this problem is fundamental: current personalization systems rely almost exclusively on behavioral proxies – clicks, time spent, purchase history – for user personalization, with few in-roads for direct expressions of user intent.[9]

While useful as a first step for user personalization, exclusive reliance on inference 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. [2024] call this 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. One way to query user 'mental states' more directly is enabling users to state them.

HCP addresses this larger ecosystem challenge by providing a mechanism for **direct preference articulation** to supplement and ground inference. For instance, I may explicitly declare a preference to gate against tabloid gossip, thereby using my HCP as a mechanism to state higher-ordered preferences difficult to infer from actions alone. Here, it is the fact that LLM reasoning can be directed via natural language that makes the consumer application low friction. 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. This is an argument for *correcting* inferred preferences via an HCP.

### 4.3 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; Poole-Dayan et al., 2025]. 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; Kleinberg and Raghavan, 2021].

HCP offers an architectural solution by letting users control their preference data and how it is used. This can counteract homogenization by supplying models with explicit, diverse preference signals. In addition, HCP supports *steerable pluralism*, where models can be guided to reflect specific user-defined viewpoints [Sorensen et al., 2024]. Importantly, this does not require fine-tuning on a single user. Instead, HCP can support opt-in sharing of specific preference categories so models can learn from groups of users (e.g., "similar users") in a way that is explicit and consented, improving data volume while keeping value diversity.

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 4.4.4.

### 4.4 Practical challenges

#### 4.4.1 Standards convergence

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]. It is our belief that academic discussion here would be particularly useful in guiding industry towards appropriate standards in the evolving economy over personalization.

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.

Yet, for these concerns brought on by a desire to discipline the market, the market may yet be the solution. In particular, if new firms can enter which provide preference management solutions superior to those provided by incumbents,

[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].

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.

#### 4.4.2 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.

#### 4.4.3 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.

#### 4.4.4 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.

We view each of the difficulties listed in this section not as insurmountable obstacles, but as research questions worthy of collaborative effort.

## 5 Use cases

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 novel 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.

### 5.1 Expanding user agency: from preference expression to discovery

Many industries across the information economy rely on preference formation and learning. While many models assume stable, well-defined preferences, evidence shows that consumers engage in costly search and experimentation to learn what they actually want [DellaVigna, 2009]. Work on experience goods highlights that preferences for many products cannot be evaluated without trial [Nelson, 1970], and research on constructive choice shows that preferences are often formed during decision-making rather than simply retrieved [Bettman et al., 1998].

HCP enables **systematic preference discovery** through controlled self-experimentation across AI systems and contexts. Instead of treating preferences as static settings, users can try alternative preference profiles, observe downstream behavior, and iteratively refine what they endorse. This has precedent in the Quantified Self movement [Swan, 2012], but HCP extends the idea from physiological tracking to values, information diets, and decision rules.

**News and information.** Users can experiment with preferences over breadth vs. depth, source diversity, topic mix, and framing. This is especially useful when users are unsure what informational style helps them (e.g., short summaries vs. long-form analysis). Users can also test different lenses for the same topic (including political lenses) to better understand their own reactions and priorities [Druckman and Lupia, 2000]. Critically, this does not have to be tied to any single domain: HCP can store multiple *time-indexed* profiles ("what I wanted at 20, what I want now") and let users compare outcomes against their own past baselines, making preference change legible rather than implicit.

**Matching markets.** Many matching markets clear based on preferences under search frictions, and better preference clarity can improve outcomes. Two salient examples are dating [Rosenfeld and Thomas, 2012; Finkel et al., 2012; Hitsch et al., 2010] and school placement (e.g., college admissions or the NRMP match) [Roth and Peranson, 1999; Gale and Shapley, 1962]. HCP can help users articulate and update constraints (dealbreakers, tradeoffs, long-run vs. short-run priorities), then test how those choices change suggested matches or application strategies, without rebuilding the preference profile from scratch each time.

**Product discovery.** Most obviously, the global digital advertising industry – worth over $600 billion annually [eMarketer/Insider Intelligence, 2024] – is built around reducing search costs and helping users discover products. HCP allows a more user-directed version of discovery: users can specify what kinds of products and pitches they want to see, what they want to avoid, and what tradeoffs they want to optimize (price, sustainability, quality, novelty). The same mechanism can cover entertainment discovery (e.g., what to watch or listen to) as a lower-stakes sandbox for testing and refining preferences.

In general, HCP's preference-discovery value is highest for *high-dimensional* preferences, in markets with substantial product diversity (where expansive search is costly), and in one-shot, high-stakes scenarios – precisely the settings where today's personalization is hardest to control.

### 5.2 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].

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. In practice, these mechanisms are typically executed by user-side assistants ("personal agents") acting on the user's behalf across services, and will only continue to do so as transaction costs continue to decrease [Shahidi et al., 2025]. 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 browser 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 go 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. 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.

**Public services.** Public-sector services are a natural fit for HCP-style preference infrastructure because they often require personalization while operating under strong constraints around consent, purpose limitation, and accountability. Across domains like education, healthcare, and benefits administration, individuals repeatedly re-specify the same constraints (needs, communication preferences, accessibility requirements, eligibility-relevant context), and each institution ends up maintaining its own partial, non-portable profile. HCP offers a user-governed way to carry these preferences across services, while still enabling fine-grained, purpose-scoped sharing and revocation.

Educational institutions could provide HCP infrastructure as a complement to the digital investments (e.g., laptops or tablets) made in schools. This would let students (and, where appropriate, guardians or school administrators) maintain a persistent set of learning-relevant preferences – such as accessibility accommodations, language support, pacing, and feedback style – and share only the relevant subset with specific tools or vendors. Different pedagogical approaches could then be implemented through how institutions and educators choose to request and apply these preferences, while students retain a clear, inspectable record of what is being shared and why. This enables personalized learning without forcing each platform to build its own closed preference silo.

**Democratic governance.** HCP could support democratic governance by enabling citizens to share *purpose-scoped* preference profiles with elected representatives or public institutions. The point is not to replace elections (or deliberative mini-publics), but to complement low-frequency, low-bandwidth voting with higher-resolution, consented signals about priorities, tradeoffs, and constraints on specific policy areas. In this sense, HCP would let governments form more fine-grained representations of public will than crude polling or whoever is loudest in public comment channels, while preserving privacy through explicit user authorization.

Concretely, HCP could enable new aggregation and interaction patterns. For example, governments could run issue-specific consultations where citizens contribute (or selectively reuse) relevant preference snippets, then apply clustering and sensemaking methods to identify stable coalitions and common-ground statements at scale (as in Polis-style opinion mapping) [Small et al., 2021; Konya et al., 2025]. Systems that use HCP can also feed deliberative processes by helping select agendas, generate structured alternatives, and surface value conflicts for deeper discussion, rather than treating deliberation as a one-off event [Fishkin and Luskin, 2005].

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

### 5.3 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.

### 5.4 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 [GDPR, 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.

## 6 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 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. 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, Xiaoxue Du, Matt Gee, Dazza Greenwood, Thomas Hardjono, Galen Hines-Pierce, Joshua Levy, and Robert Mahari for helpful conversations and feedback on earlier drafts. We also thank the Cooperative AI Foundation for their support.

## 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-soverereign-identity.html`, 2016.

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

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*, 2021.

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.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Anna Goldie, Azalia Mirhoseini, and et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

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.

Carter Blair, Kate Larson, and Edith Law. Reflective verbal reward design for pluralistic alignment. *arXiv preprint arXiv:2506.17834*, 2025.

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.

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.

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.

GDPR. 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.

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 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.

Kunal Handa, Michael Stern, Saffron Huang, Jerry Hong, Esin Durmus, Miles McCain, Grace Yun, AJ Alt, Thomas Millar, Alex Tamkin, Jane Leibrock, Stuart Ritchie, and Deep Ganguli. Introducing anthropic interviewer: What 1,250 professionals told us about working with ai, 2025. URL `https://anthropic.com/research/anthropic-interviewer`.

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.

Natasha Jaques, Asma Ghandeharioun, Judy Hanwen Shen, Craig Ferguson, Agata Lapedriza, Noah Jones, Shixiang Gu, and Rosalind Picard. Way off-policy batch deep reinforcement learning of implicit human preferences in dialog. *arXiv preprint arXiv:1907.00456*, 2019.

Natasha Jaques, Judy Hanwen Shen, Asma Ghandeharioun, Craig Ferguson, Agata Lapedriza, Noah Jones, Shixiang Gu, and Rosalind Picard. Human-centric dialog training via offline reinforcement learning. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 3985–4003, 2020.

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 and Manish Raghavan. Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences*, 118(22):e2018340118, 2021.

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, 2024.

Andrew Konya, Luke Thorburn, Wasim Almasri, Oded Adomi Leshem, Ariel Procaccia, Lisa Schirch, and Michiel Bakker. Using collective dialogues and ai to find common ground between israeli and palestinian peacebuilders. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pages 312–333, 2025.

David Laibson. Golden eggs and hyperbolic discounting. *The Quarterly Journal of Economics*, 112(2):443–478, 1997.

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

OpenAI. Model spec. `https://model-spec.openai.com/`, 2025b. Accessed: 2025-12-13.

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`.

Elinor Poole-Dayan, Jiayi Wu, Taylor Sorensen, Jiaxin Pei, and Michiel A Bakker. Benchmarking overton pluralism in llms. *arXiv preprint arXiv:2512.01351*, 2025.

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.

Peyman Shahidi, Gili Rusak, Benjamin S Manning, Andrey Fradkin, and John J Horton. The coasean singularity? demand, supply, and market design with ai agents. Working Paper 34468, National Bureau of Economic Research, November 2025. URL `http://www.nber.org/papers/w34468`.

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.

Taylor Sorensen, Pushkar Mishra, Roma Patel, Michael Henry Tessler, Michiel A Bakker, Georgina Evans, Iason Gabriel, Noah Goodman, and Verena Rieser. Value profiles for encoding human variation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2047–2095, 2025.

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*, 2025a.

Tobin South, Subramanya Nagabhushanaradhya, Ayesha Dissanayaka, Sarah Cecchetti, George Fletcher, Victor Lu, Aldo Pietropaolo, Dean H. Saxe, Jeff Lombardo, Abhishek Maligehalli Shivalingaiah, Stan Bounev, Alex Keisner, Andor Kesselman, Zack Proser, Ginny Fahs, Andrew Bunyea, Ben Moskowitz, Atul Tulshibagwale, Dazza Greenwood, Jiaxin Pei, and Alex Pentland. Identity management for agentic ai: The new frontier of authorization, authentication, and security for an ai agent world. *OpenID Foundation Whitepaper*, 2025b.

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.

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.

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.