Human Model: The Missing Piece Toward Trustworthy AGI

Zonghuan Xu[1] **Xingjun Ma**[1,†] **Yu-Gang Jiang**[1,†]
[1]Institute of Trustworthy Embodied AI, Fudan University, Shanghai, China
Shanghai Key Laboratory of Multimodal Embodied AI, Shanghai, China
[†]Corresponding authors: xingjunma@fudan.edu.cn, ygj@fudan.edu.cn

Abstract

Modern AI progress has been organized around tasks, data, and metrics that make capabilities trainable, comparable, and scalable. This organization makes external task performance easier to train and evaluate, but naturally leaves underdeveloped the question of how AI systems should understand people and use that understanding when they act. Human information already enters AI through feedback, preference data, personalization, evaluation, and deployment analysis, yet these uses often remain setting-specific or limited in depth. We introduce the *human model* concept as a unifying perspective on such work. Existing human-related AI methods can be viewed as partial forms of human-state modeling across cognition, affect, and behaviour. Under this perspective, we formulate the generalization hypothesis: whether, to what extent, and in what forms setting-specific human models can generalize. We argue that testing this generalization hypothesis should become a central research agenda for AI, and discuss what forms of data infrastructure could turn it into an empirical research problem and support the development of scalable human models. If human models can support generalizable understanding of people, they may become a crucial component toward trustworthy AGI.

1 Introduction

Since Turing asked whether machines can think, artificial intelligence has carried an ambition larger than isolated task success: to build systems whose intelligence is meaningful in relation to human thought, behaviour, and society [1]. The Dartmouth proposal gave this ambition a technical form, suggesting that central features of intelligence could be described precisely enough for machines to simulate them [2]. Along this tradition, AI research seeks to understand intelligence while also turning that pursuit into trainable, comparable, and cumulative research targets. Tasks, data, and evaluation have therefore become the basic organizing forms of modern AI progress. A task specifies what the system should do; data provide learning material; metrics allow systems to be compared. Through this organization, progress in vision, language, and agent research becomes trainable, measurable, and scalable [3, 4, 5, 6, 7, 8, 9, 10, 11]. This organization also shapes what the field can easily see and improve. Capabilities that can be expressed as tasks, datasets, and metrics are easier to train, evaluate, and scale; by contrast, it is harder to formalize how an AI system should understand people and use that understanding when it acts. This is the starting point of our paper: if the long-term aim of AI remains tied to intelligence in human worlds, then how AI systems understand people, and how they affect people, should become core questions for AI research.

The issue is methodological as well as historical. Measurement work has long emphasized that operationalizations carry assumptions about the constructs they claim to measure [12]. Recent analyses of AI benchmarks similarly show that static datasets, narrow metrics, and leaderboard incentives can shape the direction of model development [13, 14]. In AI, this produces an asymmetry between human-centered goals and task-centered research structures. Human information often enters AI research indirectly, through demonstrations, preference data, safety filtering, post-hoc evaluation,

Preprint.

or deployment analysis. This pattern appears in imitation learning, RLHF and DPO, constitutional or AI-feedback alignment, red-teaming, and human-AI evaluation [15, 16, 17, 18, 19, 20, 13, 10]. These mechanisms are important, but they usually position people as sources of labels, constraints, or outcomes. The central modeling targets remain much more often external tasks, environments, and system states. As a result, success is clearest when answers are correct, tasks are completed, or scores improve. These measures capture task outcomes much more directly than what people believe, understand or experience. Recent evaluation work already points to this gap between task performance and human experience. HELM [13] argues for metrics beyond accuracy; ChatBench [21] shows that AI-alone accuracy can fail to predict user-AI accuracy; user studies of AI chat assistants show that satisfaction, adoption, and frustration depend on usability, expectations, and interaction context, not benchmark performance alone [22].

Several lines of work have already explored more substantive ways for AI systems to use and model information about people. User modeling captures user profiles and preferences [23, 24]; knowledge tracing captures changing learner knowledge [25, 26]; and trust estimation captures trust and reliance [27, 28]. They treat aspects of people as variables to be inferred, updated, and used by AI systems, suggesting that information about people is learnable in specific settings and can improve prediction, adaptation, and system behaviour. If such models remain setting-specific, they remain isolated successes. If they generalize, they point toward a general capability for AI systems that must understand and act with people. This raises the central question of generalization.

In modern AI, a major qualitative shift often occurs when information collected in specific settings is organized into shared training problems that produce reusable representations: image labels support visual representations, next-token prediction supports broad language competence, and environment prediction supports control and planning. Information about people has not yet been organized into an analogous shared modeling problem, let alone a mature shared training objective. We use the term *human model* to name this target: computational models that estimate human states, model how they may change, and make these estimates usable for prediction, evaluation, or action selection. Here, human states refer to cognition, affect, and behaviour, from beliefs and emotions to intentions and choices. These states are shaped by roles, routines, norms, and institutional constraints, reflecting the interplay between individuals and the social structures in which they unfold. Under this concept, we formulate the generalization hypothesis: whether, to what extent, and in what forms setting-specific human models can generalize.

Our position. AI research should make the generalization hypothesis central to human model research. AGI requires both generality and trustworthiness: it must generalize across tasks and environments while remaining reliable for the people who use it, depend on it, and are affected by it. Bringing these two requirements together means that AI systems must also generalize in how they understand people, predict human behaviour and human-side consequences, and use such predictions for action selection. If human models can support this capacity across users, tasks, systems, and time, they may become a crucial component on the path toward trustworthy AGI.

Our contributions.

- **Human model as a unifying perspective.** We introduce the human model concept as a way to view scattered human-related AI methods as partial forms of human-state modeling.
- **The generalization hypothesis as the central question.** We formulate the generalization hypothesis to make explicit the key research question behind this perspective: whether, to what extent, and in what forms setting-specific human models can generalize.
- **Data infrastructure for human-model generalization.** We discuss what forms of data infrastructure could turn the generalization hypothesis into an empirical research problem and support the development of scalable human models. As one direction, we propose situated experience records as a promising data form.

Figure 1 summarizes the paper's thesis. Panel A shows the human model as a perspective that reorganizes scattered human-related AI methods into a common modeling target. Panel B formulates the generalization hypothesis as movement from setting-specific human models toward models with greater depth and scope. Panel C identifies data infrastructure as the empirical foundation needed to make this agenda testable and scalable.

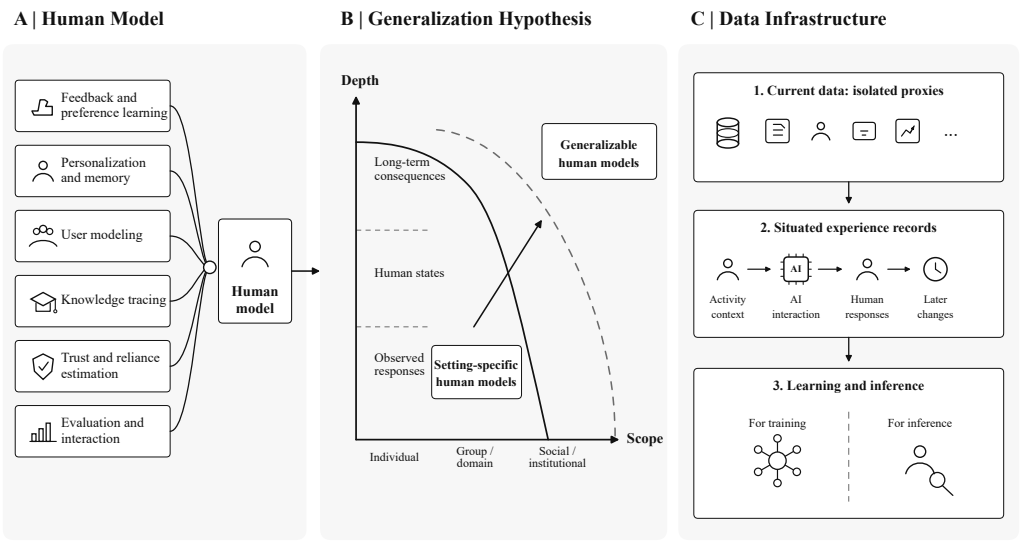

Figure 1: **The human model perspective and its research agenda.** Panel A illustrates the human model perspective: scattered human-related AI methods can be reorganized as partial, setting-specific human models rather than as isolated techniques. The generalization hypothesis asks whether, to what extent, and in what forms such setting-specific human models can generalize. In panel B, both curves are schematic and do not represent precise empirical boundaries: the solid curve indicates the limited depth and scope of current setting-specific human models, while the dashed curve indicates the hypothesized region of generalizable human models. Panel C abstracts the needed data infrastructure: isolated proxies are connected into situated experience records, making human-model learning, inference, and evaluation empirically testable.

2 Existing Work Through the Human Model Perspective

Section 1 presented the human model concept as a unifying perspective on existing human-related AI methods. We now use this perspective to organize representative lines of work. At first glance, these lines may seem only loosely related. Viewed through the human model perspective, however, they can be organized by asking what aspect of human states they model and how that information is used by AI systems.

Practical use. The rows in Table 1 already affect AI systems in concrete ways. RLHF, DPO, and reward models change language models through human judgments. Memory, personalization, and profiles change how systems respond to particular users over time. Knowledge tracing uses learner-state estimates to predict future performance and support adaptive tutoring [25, 26, 49]. Trust and reliance studies show that human confidence in automation affects use, misuse, and overreliance [28, 50, 51, 52]. In HRI, modeling trust and adaptation can change robot behaviour and collaboration outcomes [73, 61]. These examples show that information about people carries predictive structure and can change system behaviour, evaluation, and adaptation.

Toward transfer. Several of these lines already point toward generalization. Knowledge tracing studies transfer across students, concepts, and platforms [49]. User modeling and personalization increasingly aim at reusable user representations across tasks and long-term contexts [42, 71]. HRI and trust-aware planning have moved from single trust variables toward richer online trust and collaboration models [73, 61, 74]. These efforts motivate our generalization hypothesis.

3 The Generalization Hypothesis

Section 1 introduced the generalization hypothesis. Section 2 showed why the generalization hypothesis is not merely speculative: existing work already shows that information about people

Table 1: Existing human-related AI methods viewed through the human model perspective. Each row can be read as a partial human model, differing in what human information it captures and how that information enters AI systems.

Line of work	What it captures about people	Human-model role
RLHF / DPO / reward models	Evaluative judgments from feedback, preferences, and ratings [15, 29, 16, 17, 30, 31]	Learning signal for modeling evaluative judgment.
Preference alignment	User or group preferences over system behaviours [18, 19, 32, 33, 34]	Target specification for preference-sensitive behaviour.
Reward inference / assistance games	Latent objectives inferred from behaviour or comparisons [35, 36, 32]	Bridge from objective inference to assistance and action selection.
Personalization / memory	Persistent user-specific states across interactions [37, 38, 39, 40]	User-level state for adaptation over time.
User modeling / profiles / persona	Traits, contexts, roles, and identity patterns of users [23, 24, 41, 42, 43]	Classical basis for user-level representations.
Theory of mind / intent recognition	Beliefs, intentions, plans, and other latent mental states behind behaviour [44, 45, 46, 47]	Cognitive basis for latent-state inference.
Knowledge tracing / learner modeling	Learner knowledge, misconceptions, and update dynamics over time [25, 26, 48, 49]	Domain example of state-change modeling over time.
Trust estimation / reliance	Trust, reliance, over-reliance, and calibration in response to automation [27, 28, 50, 51, 52]	Affective and reliance-related state for calibration.
HRI / collaboration	Human roles, intentions, constraints, and coordination states in joint activity [53, 54, 55, 56]	Interaction setting where human-state estimates guide joint action.
Human-aware planning / POMDP	Partially observed human states for action selection [57, 58, 59, 60, 61]	Decision framework that makes human-state estimates action-relevant.
Action execution / shared autonomy	User intent, capability, and control preferences during execution [62, 63, 64, 65]	Execution-time signal for adaptive assistance.
Skills / workflow customization / harnesses	Routines, tool-use patterns, task practices, and organizational workflows [66, 67, 68, 8]	Extension from individuals to situated work practices.
Human-AI evaluation	Human responses, judgments, and joint outcomes under system interaction [13, 11, 21, 22]	Evaluation setting for human responses and joint outcomes.
User studies / behavioural logs	Observed behaviour, experience, and state changes during interaction [69, 51, 70, 71, 72]	Empirical source for learning and validating human models.

can be learned, used by AI systems, and in some cases transferred beyond its original setting [75, 76, 77]. However, these signs do not yet amount to a clear research problem. To study human model generalization systematically, we need to unpack what generalization means, not treat it as a single yes-or-no outcome. This section does so in two ways. First, we ask what levels of depth and scope human models may reach. Second, we ask what technical forms such generalization may take, including the shared learning problems, prediction targets, and verifiable signals through which it could appear.

The empirical status of the generalization hypothesis should be established through concrete studies. It may turn out that human-model generalization appears only locally, in modular or domain-conditioned forms, with scaling trends limited to particular populations or contexts. It may also reveal broader and more promising regularities. In either case, mapping where generalization holds and where it fails is itself a central research objective.

Depth and scope. Table 2 summarizes the levels that human models may reach. We describe these levels along two dimensions. The first dimension is the depth of understanding people. At the shallowest level, a model captures observable responses, such as language, clicks, behaviour, facial expressions, or physiological signals. At a deeper level, it infers human states, such as beliefs, knowledge, preferences, trust, capability, or intentions. Deeper still, it predicts how AI actions may

Table 2: Depth and scope of human model generalization. The table organizes possible human models by how deeply they model people and how broadly they apply.

Scope / depth	Observed responses	Inferred human states	Changes after AI actions	Long-term outcomes
Task / setting-specific	**Individual signals:** language, clicks, choices, gaze	**Individual state:** beliefs, knowledge, preferences, trust	**Setting-specific change:** understanding, reliance, choices	**Setting-specific outcomes:** preference shift, capability change, habits
Cross-context	**Interaction signals:** turn-taking, feedback, coordination	**Multi-person state:** shared understanding, roles, team trust	**Cross-context change:** coordination, motivation, task allocation	**Cross-context outcomes:** role adaptation, workflow change, group outcomes
General / societal-level	**Aggregate signals:** population patterns, discourse, logs	**Social state:** norms, conventions, institutional constraints	**Societal change:** adoption, incentives, public response	**Societal outcomes:** norm formation, inequality, social trust

change human states. At the deepest level, it estimates longer-term human consequences, such as dependence, capability change, norm formation, or social impact. The second dimension is scope. A human model may focus on a single user or one human-AI interaction, and it may also extend to multi-person collaboration, teams, organizations, institutions, and broader social structures. These levels make human model generalization more concrete. They describe how deeply AI systems model people, how broadly these models apply, and whether they can become reusable across users, tasks, systems, and time.

Forms of generalization. Human model generalization should not be understood as a model becoming stronger on all human-related tasks at once. In modern AI, generalization usually appears through a more specific technical form: a shared learning problem, a prediction target, and a loss. In language models, the shared problem is next-token prediction, and generalization first appears as lower perplexity on held-out text [78]. In vision-language models, the shared problem is image-text prediction or matching, and generalization appears as visual representations that transfer across tasks [79]. In vision-language-action models, the shared problem is action prediction from robot demonstrations and vision-language-action data, and generalization appears as embodied policies that adapt beyond the original demonstrations [80]. Later capabilities are built on these forms, and some tasks show corresponding scaling trends while others may not.

By analogy, human model generalization is more likely to first appear through a pretraining-like shared prediction problem than through direct success on every downstream human-facing task. For human models, a plausible shared problem is to learn from records of how people encounter AI systems, how the systems act, how people respond, and what later changes follow. Testing the generalization hypothesis therefore requires systematic evaluation: where prediction and transfer improve with scale, where they fail, and what practical conditions make the route viable. The immediate bottleneck is data, which we discuss in Section 4.

4 Data Infrastructure for Human Models

Modern AI progress has repeatedly depended on data regimes that turn capabilities into trainable, scalable, and testable objectives, as shown by recent work on data-centric AI, language-model training corpora, and code benchmarks [81, 82, 83, 8]. For human models, the corresponding bottleneck is data infrastructure: what kinds of human-related data can make human states learnable, transferable, and evaluable at scale? Existing human-related data are useful but fragmented. Preference labels, user profiles, interaction logs, surveys, behavioural logs, and physiological signals are typically collected for particular tasks, products, or studies [37, 84, 85]. Such data can support setting-specific modeling, but they are rarely organized to test the generalization of human-state representations. This section therefore focuses on the data forms needed to turn human-state modeling into a scalable empirical problem. We argue that one promising direction is to move from isolated proxies toward situated experience records.

From Isolated Proxies to Situated Experience Records. Existing data forms have structural limits for studying human-model generalization. Interaction logs connect system actions with user responses, but they only partially reveal user understanding, trust, or difficulty. Speech and language are information-rich, but their meaning becomes unstable when separated from task context, what the user saw, and later outcomes. Gaze, physiological signals, and neural signals can be temporally dense,

but their connection to higher-level human states, such as understanding or intention, is indirect and noisy. Surveys and self-reports provide subjective anchors, but they are sparse, retrospective, and constrained by limits of self-knowledge and recall. Social and workflow records preserve roles, routines, and institutional context, but they often miss a person's immediate response to a specific event. Moreover, these data forms are valuable, but they usually reflect only one aspect of human state [86, 87, 85]: data collected for a specific task are formatted to make that task tractable, so signals judged irrelevant to the task are compressed into labels, treated as noise, or excluded to control collection cost. Models may then learn the structure selected by the task and its collection pipeline rather than transferable human-state structure.

This limitation is common in practice. A user accepting a suggestion may indicate trust, or it may be a default choice under time pressure. Repeatedly viewing the same content may indicate interest, confusion, or uncertainty. A positive verbal response may express genuine satisfaction, or it may serve to close the current task. Personal assistants, educational systems, workflow agents, embodied assistants, and AI evaluation settings all face this ambiguity. A more suitable organization should therefore start from the user's experience and keep relevant human-related observations together within the same activity. We call these data units *situated experience records*. They may include task background, user situation, AI actions, immediate human responses, explicit feedback, later revisions, and downstream outcomes. In such records, the human state remains a latent quantity to be inferred and validated through relations among multiple proxies. No single proxy then has to carry the full explanatory burden. Instead, multiple signals within one experience can constrain and calibrate one another, helping models learn more stable human-state structure.

Language-model pretraining provides a useful analogy [81, 83, 82]. Next-token prediction is also a proxy objective: it does not directly annotate knowledge, reasoning ability, or task competence. Yet natural language preserves many relations among knowledge, goals, task examples, communicative intent, and social context. With sufficient scale, surface token prediction can therefore support transferable representations. Similarly, human models should not expect any single observable proxy to equal a human state. Proxies such as a click, rating, pause, or acceptance is not itself trust, understanding, or intention. If records keep task context, AI action, human response, explicit feedback, and later outcome within the same experience, however, relations among proxies may provide learnable structure.

Capturing Data in the Flow of Activity. Situated experience records need to be constructed during or near ordinary activity, since many important signals naturally appear only as activity unfolds [88, 89, 90]. Traditional user studies, retrospective surveys, and product logs each provide partial information, but they are limited in scale, temporal continuity, or coverage of ordinary activity. A more promising route is to record relevant links in a natural and scalable way as activity unfolds: what the user sees, hears, does, says, accepts, corrects, and later returns to, rather than reconstructing these relations afterward through costly interviews, surveys, or isolated labels.

Recent work on computer-use and GUI-navigation agents suggests a shift toward device- and interface-level interaction. AI interfaces are moving beyond chat windows into operating systems, applications, devices, and everyday workflows. For example, OSWorld evaluates multimodal agents in real computer environments across web and desktop applications, while OSUniverse focuses on complex desktop-oriented GUI navigation tasks [91, 92]. These systems are only early signs; complete data infrastructure remains open. Their relevance is that future AI systems may be positioned closer to the activity streams from which situated experience records can be constructed.

Learning from Fine-Grained Human Responses. The training value of situated experience records lies in placing fine-grained human responses inside the structure of the same experience. A gaze shift, a pause, an acceptance, or an undo action cannot by itself be mapped directly to a human state. When combined with task context, AI interactions, explicit feedback, and later outcomes, however, these responses can become dense proxies for learning human states [86, 87].

Human-model training should therefore not rely only on explicitly elicited targets such as satisfaction scores, preference choices, or human ratings. Explicit feedback and self-reports can serve as sparse anchors, while naturally occurring fine-grained responses provide a more continuous training source. A model can then learn which signals recur across similar experiences and which vary with AI behaviour, task conditions, or later feedback. Training thus can shift from predicting a single proxy

to learning relations among multiple signals within the same experience, therefore increasing the possibility of generalization.

Using Context at Inference Time. Situated experience records also matter at inference time. Many failures of AI assistance arise not only because models lack reasoning ability, but because AI is absent from the context in which work is generated. In real work, the task itself is often short, but the context is long; the execution step may be light, but locating context, handing off state, specifying constraints, and synchronizing systems can be heavy. A request such as "forward this," "check this," "confirm this," or "remind me" may depend on messages, calls, document state, spreadsheet fields, system permissions, organizational procedures, personal memory, and local experience. If users have to reorganize this background into a prompt, the cost of handing the task to AI may exceed the cost of completing it manually.

This problem is more fundamental for deployed human models. For ordinary assistants, missing context is first a source of usability friction. For human models, missing context becomes a measurement bottleneck. A human model aims to infer latent states such as understanding, trust, confusion, intention, pressure, preference, and dependence. These states are rarely fully reportable in the user's current input. Their evidence is often distributed through the activity itself: commitments made in meetings, tone in calls, information the user saw, waiting and hesitation, later revisions, undo actions, repeated confirmations, or longer-term reliance. If such information can only be recovered through retrospective recall and manual entry, the model receives sparse, costly, and retrospectively biased self-reports rather than situated evidence from the activity.

Situated experience records can therefore serve as inference-time context infrastructure. Rather than waiting for users to package a task as a complete input, they preserve relevant task state such as information environment, interaction history, AI actions, human responses, and downstream outcomes as activity unfolds. When a human model interprets a new request, feedback, revision, or behavioural trace, it can place that proxy back into the concrete activity instead of relying on incomplete fragments actively supplied by the user. In this sense, situated experience records point toward a persistent layer: within appropriate boundaries, AI systems can carry context which is automatically generated in workflows, reducing the friction of context handoff while giving human-state inference more stable measurement conditions.

Treating Real-World Activity as Noisy Measurement. Data collected closer to real activity can also introduce stronger measurement confounding. User behaviour may be shaped by interface design, task difficulty, or environmental factors. Large-scale language and multimodal data practices further show that low-quality samples, duplicates, contamination, and misalignment can weaken generalization [83, 93, 94]. Human-model data infrastructure should therefore treat situated experience records as noisy observations. At a minimum, data infrastructure should preserve metadata, such as provenance, collection context, and generation conditions. This allows models to better distinguish transferable human-state structure from artifacts introduced by data collection.

5 Alternative View: Human Models as Part of World Models

One natural alternative view is that human models need not be treated separately: humans are part of the world, so sufficiently complete world models should eventually represent human cognition, affect, and behaviour. We agree this viewpoint ontologically, but our claim is more of methodological. Distinguishing human models makes the problem visible as a research direction in its own right, where progress can be organized around the data, objectives, and evaluations needed to model human states. Without such organization, changes in human states are easily hidden inside generic prediction or post hoc evaluation.

This distinction also matters for governance. Human-facing AI does not only predict external environments; it also affects how people understand, rely on, resist, adapt to, and are changed by AI actions. These effects raise normative and dual-use questions that are easy to obscure when human states are treated as just another part of the environment. Human models are therefore best understood as a focused part of world modeling whose data, evaluations, and risks require explicit treatment. The distinction is analogous to treating vision, language, reward modeling, or safety evaluation as distinct research directions even though all are, in principle, parts of modeling the world.

6 Discussion: Implications and Constraints

The human model perspective reorganizes human-related information that is currently scattered across AI systems. Such information is often used for specific purposes such as personalization, interface improvement, feedback training, or post-hoc evaluation. Under the human model perspective, it can also be viewed as material for learning human states and for organizing a more unified research problem around cognition, affect, and behaviour. If the generalization hypothesis holds in some form, human models may bring new system capabilities: AI systems may better predict human responses, estimate how AI actions affect people, and adjust their behaviour accordingly. Since these capabilities are directed toward people, they also raise clearer risks and stronger governance requirements.

Potential applications. Figure 2 summarizes three application patterns. First, validated human models could make human-centered evaluation more scalable: observed human signals remain necessary, but model-mediated proxies could extend human-centered signals from evaluation into training and model selection across systems, tasks, populations, and time. Second, human models could support simulation by estimating how people may respond under changing contexts, interventions, information conditions, and scenarios. Third, human models could become part of human-aware agent loops, complementing world models with estimates of how actions may affect human cognition, affect, behaviour, and potential downstream consequences.

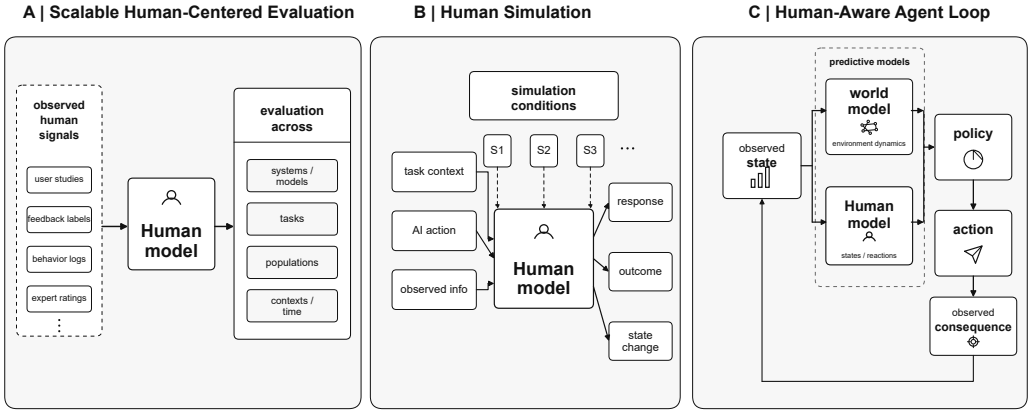

Figure 2: **Potential applications of human models.** Panel A illustrates scalable human-centered evaluation: sparse or incomplete observed human signals can calibrate a human model, which then extends evaluation across systems, tasks, populations, contexts, and time. Panel B illustrates human simulation: simulation conditions are routed through a human model to project responses, outcomes, and state changes. Panel C illustrates a human-aware agent loop: world and human models act as parallel predictive models for policy, so action selection can account for both external dynamics and human-side consequences.

Power, dual use, and data boundaries. Looking further ahead, situated experience records also point toward a stronger AI presence. If users, under explicit permission, lawful process, and clear boundaries, allow devices, operating systems, or workflow interfaces to preserve relevant activity, then meetings, calls, documents, system actions, and everyday interactions could become contextual sources for human models. Such models would not only respond to isolated prompts. They could gradually model personal states, relationships, organizational routines, and collective dynamics, and could support low-friction delegation within appropriate boundaries. A stronger human model could help an AI system detect confusion, calibrate trust, preserve user control, and adapt assistance to human needs. The same capability could also predict when a person is likely to defer, accept a suggestion, or reveal vulnerability, and may reveal patterns that users do not explicitly express, cannot easily recall, or cannot reliably self-report.

For this reason, this route raises governance challenges beyond conventional privacy. Situated records often contain not only data about one person, but also bystanders' speech and actions, relationship structures, organizational processes, and collective contexts. Highly sensitive modalities

such as speech, gaze, physiological signals, EEG, neural-interface traces, or long-term behavioural trajectories require stricter review. If such records and strong human models become concentrated in a small number of platforms or companies, they could become unprecedented sources of behavioural and social intelligence, supporting more precise influence, manipulation, dependency formation, and concentration of power. Human-model data infrastructure therefore cannot focus only on data availability. It must treat consent, purpose limitation, data minimization, access control, retention limits, deletion rights, auditability, bystander consent, and institutional concentration as central issues from the start. Privacy-preserving learning methods such as differential privacy and federated learning can support this agenda when raw human data should not be centrally exposed [95, 96]. Human models should ultimately serve human understanding, agency, and well-being, and avoid treating more effective prediction or influence as an end in itself.

7 Conclusion

This position paper argues that trustworthy AGI requires a systematic way to model people. We introduced the human model concept as a unifying perspective on learnable representations of human states across cognition, affect, and behaviour. Human models need not correspond to a single architecture: they may be explicit or implicit, setting-specific or more general. Their common role is to estimate human states, model how those states may be changed by AI actions, and make these estimates usable for prediction, evaluation, or action selection.

The central question is whether setting-specific human models can generalize. Existing work in RLHF, personalization, user modeling, knowledge tracing, trust estimation, HRI, and human-AI evaluation already provides many local forms of human-state modeling. Yet these lines have not been organized as a unified research agenda. We formulate the generalization hypothesis to move the question from whether people can be modeled in a given setting to which human-state representations can transfer across users, tasks, systems, and time. Whether such generalization turns out to be broad, local, or conditional, systematically testing its boundaries should become an important objective for AI research.

To make this problem trainable and evaluable, human models also require corresponding data infrastructure. We propose situated experience records as one promising direction: preserving task state, AI actions, human responses, explicit feedback, and downstream outcomes as activity unfolds, so that models can learn human states from relations among multiple proxies within the same experience. Such records are not only useful for training. At inference time, they can form a persistent context layer, reducing the friction of repeatedly handing off context and giving human-state inference more stable measurement conditions. This route remains only a starting point, and its technical, ethical, safety, and governance boundaries require further study. But if the generalization hypothesis holds in some form, human models are likely to become an important component of trustworthy AGI: AI systems that generalize over external tasks and environments, and also develop transferable ways to understand people.

References

[1] A. M. Turing. Computing Machinery and Intelligence. *Mind*, LIX(236):433–460, 1950. doi: 10.1093/mind/LIX.236.433. URL https://academic.oup.com/mind/article/LIX/236/433/986238.

[2] John McCarthy, Marvin L. Minsky, Nathaniel Rochester, and Claude E. Shannon. A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 1955. URL https://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html. Dated August 31, 1955.

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE Computer Society, 2009. doi: 10.1109/CVPR.2009.5206848. URL https://doi.org/10.1109/CVPR.2009.5206848.

[4] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding, 2019. URL https://arxiv.org/abs/1804.07461.

[5] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2021. URL https://arxiv.org/abs/2009.03300.

[6] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021. URL https://arxiv.org/abs/2110.14168.

[7] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code, 2021. URL https://arxiv.org/abs/2107.03374.

[8] Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. Swe-bench: Can language models resolve real-world github issues?, 2024. URL https://arxiv.org/abs/2310.06770.

[9] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. Gpqa: A graduate-level google-proof q&a benchmark, 2023. URL https://arxiv.org/abs/2311.12022.

[10] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023. URL https://arxiv.org/abs/2306.05685.

[11] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating llms by human preference, 2024. URL https://arxiv.org/abs/2403.04132.

[12] Abigail Z. Jacobs and Hanna Wallach. Measurement and fairness, 2021. URL https://arxiv.org/abs/1912.05511.

[13] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic evaluation of language models, 2023. URL https://arxiv.org/abs/2211.09110.

[14] Maria Eriksson, Erasmo Purificato, Arman Noroozian, Joao Vinagre, Guillaume Chaslot, Emilia Gomez, and David Fernandez-Llorca. Can we trust ai benchmarks? an interdisciplinary review of current issues in ai evaluation, 2025. URL https://arxiv.org/abs/2502.06559.

[15] Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *arXiv preprint arXiv:1706.03741*, 2017. URL https://arxiv.org/abs/1706.03741.

[16] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022. URL https://arxiv.org/abs/2203.02155.

[17] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. *arXiv preprint arXiv:2305.18290*, 2023. URL https://arxiv.org/abs/2305.18290.

[18] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional AI: Harmlessness from AI Feedback. *arXiv preprint arXiv:2212.08073*, 2022. URL https://arxiv.org/abs/2212.08073.

[19] Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. RLAIF vs. RLHF: Scaling Reinforcement Learning from Human Feedback with AI Feedback. *arXiv preprint arXiv:2309.00267*, 2023. URL https://arxiv.org/abs/2309.00267.

[20] Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, Andy Jones, Sam Bowman, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Nelson Elhage, Sheer El-Showk, Stanislav Fort, Zac Hatfield-Dodds, Tom Henighan, Danny Hernandez, Tristan Hume, Josh Jacobson, Scott Johnston, Shauna Kravec, Catherine Olsson, Sam Ringer, Eli Tran-Johnson, Dario Amodei, Tom Brown, Nicholas Joseph, Sam McCandlish, Chris Olah, Jared Kaplan, and Jack Clark. Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned. *arXiv preprint arXiv:2209.07858*, 2022. URL https://arxiv.org/abs/2209.07858.

[21] Serina Chang, Ashton Anderson, and Jake M. Hofman. Chatbench: From static benchmarks to human-ai evaluation, 2025. URL https://arxiv.org/abs/2504.07114.

[22] Moiz Sadiq Awan, Muhammad Haris Noor, and Muhammad Salman Munaf. Beyond benchmarks: How users evaluate ai chat assistants, 2026. URL https://arxiv.org/abs/2603.25220.

[23] Elaine Rich. User Modeling via Stereotypes. *Cognitive Science*, 3(4):329–354, 1979. doi: 10.1016/S0364-0213(79)80012-9. URL https://www.sciencedirect.com/journal/cognitive-science/vol/3/issue/4.

[24] Alfred Kobsa and Wolfgang Wahlster, editors. *User Models in Dialog Systems*. Symbolic Computation. Springer, Berlin, Heidelberg, 1989. doi: 10.1007/978-3-642-83230-7. URL https://link.springer.com/book/10.1007/978-3-642-83230-7.

[25] Albert T. Corbett and John R. Anderson. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modelling and User-Adapted Interaction*, 4(4):253–278, 1995. doi: 10.1007/bf01099821. URL https://doi.org/10.1007/bf01099821.

[26] Chris Piech, Jonathan Spencer, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas Guibas, and Jascha Sohl-Dickstein. Deep Knowledge Tracing. *arXiv preprint arXiv:1506.05908*, 2015. URL https://arxiv.org/abs/1506.05908.

[27] Bonnie M. Muir. Trust between Humans and Machines, and the Design of Decision Aids. *International Journal of Man-Machine Studies*, 27(5–6):527–539, 1987. doi: 10.1016/S0020-7373(87)80013-5.

[28] J. D. Lee and K. A. See. Trust in Automation: Designing for Appropriate Reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 46(1):50–80, 2004. doi: 10.1518/hfes.46.1.50_30392. URL https://doi.org/10.1518/hfes.46.1.50_30392.

[29] Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-Tuning Language Models from Human Preferences. *arXiv preprint arXiv:1909.08593*, 2019. URL https://arxiv.org/abs/1909.08593.

[30] Timo Kaufmann, Paul Weng, Viktor Bengs, and Eyke Hullermeier. A Survey of Reinforcement Learning from Human Feedback. *arXiv preprint arXiv:2312.14925*, 2023. URL https://arxiv.org/abs/2312.14925.

[31] 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.

[32] Siddharth Reddy, Anca Dragan, Sergey Levine, Shane Legg, and Jan Leike. Learning Human Objectives by Evaluating Hypothetical Behavior. In *International Conference on Machine Learning*, pages 8020–8029. PMLR, 2020. URL https://proceedings.mlr.press/v119/reddy20a.html.

[33] Genta Indra Winata, Hanyang Zhao, Anirban Das, Wenpin Tang, David D. Yao, Shi-Xiong Zhang, and Sambit Sahu. Preference Tuning with Human Feedback on Language, Speech, and Vision Tasks: A Survey. *arXiv preprint arXiv:2409.11564*, 2024. URL https://arxiv.org/abs/2409.11564.

[34] Daiwei Chen, Yi Chen, Aniket Rege, Zhi Wang, and Ramya Korlakai Vinayak. PAL: Sample-Efficient Personalized Reward Modeling for Pluralistic Alignment. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=1kFDrYCuSu.

[35] Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, and Stuart Russell. Cooperative Inverse Reinforcement Learning. *arXiv preprint arXiv:1606.03137*, 2016. URL https://arxiv.org/abs/1606.03137.

[36] Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, and Stuart Russell. The Off-Switch Game, 2017. URL https://www.ijcai.org/Proceedings/2017/32.

[37] Zhehao Zhang, Ryan A. Rossi, Branislav Kveton, Yijia Shao, Diyi Yang, Hamed Zamani, Franck Dernoncourt, Joe Barrow, Tong Yu, Sungchul Kim, Ruiyi Zhang, Jiuxiang Gu, Tyler Derr, Hongjie Chen, Junda Wu, Xiang Chen, Zichao Wang, Subrata Mitra, Nedim Lipka, Nesreen Ahmed, and Yu Wang. Personalization of Large Language Models: A Survey. *arXiv preprint arXiv:2411.00027*, 2024. URL https://arxiv.org/abs/2411.00027.

[38] Jiahong Liu, Zexuan Qiu, Zhongyang Li, Quanyu Dai, Wenhao Yu, Jieming Zhu, Minda Hu, Menglin Yang, Tat-Seng Chua, and Irwin King. A Survey of Personalized Large Language Models: Progress and Future Directions. *arXiv preprint arXiv:2502.11528*, 2025. URL https://arxiv.org/abs/2502.11528.

[39] Yuyang Hu, Shichun Liu, Yanwei Yue, Guibin Zhang, Boyang Liu, Fangyi Zhu, Jiahang Lin, Honglin Guo, Shihan Dou, Zhiheng Xi, Senjie Jin, Jiejun Tan, Yanbin Yin, Jiongnan Liu, Zeyu Zhang, Zhongxiang Sun, Yutao Zhu, Hao Sun, Boci Peng, Zhenrong Cheng, Xuanbo Fan, Jiaxin Guo, Xinlei Yu, Zhenhong Zhou, Zewen Hu, Jiahao Huo, Junhao Wang, Yuwei Niu, Yu Wang, Zhenfei Yin, Xiaobin Hu, Yue Liao, Qiankun Li, Kun Wang, Wangchunshu Zhou, Yixin Liu, Dawei Cheng, Qi Zhang, Tao Gui, Shirui Pan, Yan Zhang, Philip Torr, Zhicheng Dou, Ji-Rong Wen, Xuanjing Huang, Yu-Gang Jiang, and Shuicheng Yan. Memory in the Age of AI Agents, 2026. URL https://arxiv.org/abs/2512.13564.

[40] Bowen Jiang, Yuan Yuan, Maohao Shen, Zhuoqun Hao, Zhangchen Xu, Zichen Chen, Ziyi Liu, Anvesh Rao Vijjini, Jiashu He, Hanchao Yu, Radha Poovendran, Gregory Wornell, Lyle Ungar, Dan Roth, Sihao Chen, and Camillo Jose Taylor. PersonaMem-v2: Towards Personalized Intelligence via Learning Implicit User Personas and Agentic Memory, 2025. URL https://arxiv.org/abs/2512.06688.

[41] Geoffrey I. Webb, Michael J. Pazzani, and Daniel Billsus. Machine Learning for User Modeling. *User Modeling and User-Adapted Interaction*, 11(1-2):19–29, 2001. doi: 10.1023/a:1011117102175. URL https://doi.org/10.1023/a:1011117102175.

[42] Erasmo Purificato, Ludovico Boratto, and Ernesto William De Luca. User Modeling and User Profiling: A Comprehensive Survey. *arXiv preprint arXiv:2402.09660*, 2024. URL https://arxiv.org/abs/2402.09660.

[43] Jiwei Li, Michel Galley, Chris Brockett, Georgios P. Spithourakis, Jianfeng Gao, and Bill Dolan. A Persona-Based Neural Conversation Model. *arXiv preprint arXiv:1603.06155*, 2016. URL https://arxiv.org/abs/1603.06155.

[44] David Premack and Guy Woodruff. Does the Chimpanzee Have a Theory of Mind? *Behavioral and Brain Sciences*, 1(4):515–526, 1978. doi: 10.1017/S0140525X00076512. URL https://www.cambridge.org/core/journals/behavioral-and-brain-sciences/article/does-the-chimpanzee-have-a-theory-of-mind/1E96B02CD9850016B7C93BC6D2FEF1D0.

[45] Neil C. Rabinowitz, Frank Perbet, H. Francis Song, Chiyuan Zhang, S. M. Ali Eslami, and Matthew Botvinick. Machine Theory of Mind. *arXiv preprint arXiv:1802.07740*, 2018. URL https://arxiv.org/abs/1802.07740.

[46] Henry A. Kautz and James F. Allen. Generalized Plan Recognition. In *Proceedings of the Fifth National Conference on Artificial Intelligence*, pages 32–37. AAAI Press, 1986. URL https://archive.aaai.org/Library/AAAI/1986/aaai86-006.php.

[47] Gita Sukthankar, Christopher Geib, Hung Hai Bui, David V. Pynadath, and Robert P. Goldman, editors. *Plan, Activity, and Intent Recognition: Theory and Practice*. Elsevier, 2014. URL https://shop.elsevier.com/books/plan-activity-and-intent-recognition/sukthankar/978-0-12-398532-3.

[48] Benjamin Valdes Aguirre, Jorge A. Ramirez Uresti, and Benedict du Boulay. An Analysis of Student Model Portability. *International Journal of Artificial Intelligence in Education*, 26 (3):932–974, 2016. doi: 10.1007/s40593-016-0113-0. URL https://doi.org/10.1007/s40593-016-0113-0.

[49] Shuanghong Shen, Qi Liu, Zhenya Huang, Yonghe Zheng, Minghao Yin, Minjuan Wang, and Enhong Chen. A Survey of Knowledge Tracing: Models, Variants, and Applications. *IEEE Transactions on Learning Technologies*, 17:1858–1879, 2024. doi: 10.1109/tlt.2024.3383325. URL https://doi.org/10.1109/tlt.2024.3383325.

[50] Kevin Anthony Hoff and Masooda Bashir. Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 57(3):407–434, 2015. doi: 10.1177/0018720814547570. URL https://doi.org/10.1177/0018720814547570.

[51] Zana Bucinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–21, 2021. doi: 10.1145/3449287. URL https://doi.org/10.1145/3449287.

[52] Max Schemmer, Niklas Kuehl, Carina Benz, Andrea Bartos, and Gerhard Satzger. Appropriate Reliance on AI Advice: Conceptualization and the Effect of Explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, pages 410–422. ACM, 2023. doi: 10.1145/3581641.3584066. URL https://doi.org/10.1145/3581641.3584066.

[53] Thomas B. Sheridan, William L. Verplank, and Thomas L. Brooks. Human/Computer Control of Undersea Teleoperators. Technical report, Massachusetts Institute of Technology, Man-Machine Systems Laboratory, 1978. URL https://ntrs.nasa.gov/citations/19790007441.

[54] Stefanos Nikolaidis, Przemyslaw Lasota, Ramya Ramakrishnan, and Julie Shah. Improved human–robot team performance through cross-training, an approach inspired by human team training practices. *The International Journal of Robotics Research*, 34(14):1711–1730, 2015. doi: 10.1177/0278364915609673. URL https://doi.org/10.1177/0278364915609673.

[55] Stefanos Nikolaidis, David Hsu, and Siddhartha Srinivasa. Human-robot mutual adaptation in collaborative tasks: Models and experiments. *The International Journal of Robotics Research*, 36(5-7):618–634, 2017. doi: 10.1177/0278364917690593. URL https://doi.org/10.1177/0278364917690593.

[56] Samuel Westby and Christoph Riedl. Collective Intelligence in Human-AI Teams: A Bayesian Theory of Mind Approach. *arXiv preprint arXiv:2208.11660*, 2022. URL https://arxiv.org/abs/2208.11660.

[57] Sandra Carberry. Techniques for Plan Recognition. *User Modeling and User-Adapted Interaction*, 11(1-2):31–48, 2001. doi: 10.1023/a:1011118925938. URL https://doi.org/10.1023/a:1011118925938.

[58] Steve Young, Milica Gasic, Blaise Thomson, and Jason D. Williams. POMDP-Based Statistical Spoken Dialog Systems: A Review. *Proceedings of the IEEE*, 101(5):1160–1179, 2013. doi: 10.1109/jproc.2012.2225812. URL https://doi.org/10.1109/jproc.2012.2225812.

[59] Yu Zhang, Sarath Sreedharan, Anagha Kulkarni, Tathagata Chakraborti, Hankz Hankui Zhuo, and Subbarao Kambhampati. Plan Explicability and Predictability for Robot Task Planning. *arXiv preprint arXiv:1511.08158*, 2015. URL https://arxiv.org/abs/1511.08158.

[60] Tathagata Chakraborti, Sarath Sreedharan, and Subbarao Kambhampati. Balancing Explicability and Explanation in Human-Aware Planning. *arXiv preprint arXiv:1708.00543*, 2017. URL https://arxiv.org/abs/1708.00543.

[61] Min Chen, Stefanos Nikolaidis, Harold Soh, David Hsu, and Siddhartha Srinivasa. Trust-Aware Decision Making for Human-Robot Collaboration: Model Learning and Planning. *arXiv preprint arXiv:1801.04099*, 2018. URL https://arxiv.org/abs/1801.04099.

[62] Anca Dragan and Siddhartha Srinivasa. Generating Legible Motion. In *Robotics: Science and Systems IX*, volume 09, 2013. URL https://www.roboticsproceedings.org/rss09/p24.html.

[63] Shervin Javdani, Siddhartha S. Srinivasa, and J. Andrew Bagnell. Shared Autonomy via Hindsight Optimization. *arXiv preprint arXiv:1503.07619*, 2015. URL https://arxiv.org/abs/1503.07619.

[64] Jae Sung Park, Chonhyon Park, and Dinesh Manocha. I-Planner: Intention-Aware Motion Planning Using Learning Based Human Motion Prediction. *arXiv preprint arXiv:1608.04837*, 2016. URL https://arxiv.org/abs/1608.04837.

[65] Siddharth Reddy, Anca D. Dragan, and Sergey Levine. Shared Autonomy via Deep Reinforcement Learning. *arXiv preprint arXiv:1802.01744*, 2018. URL https://arxiv.org/abs/1802.01744.

[66] Lucy A. Suchman. *Plans and Situated Actions: The Problem of Human-Machine Communication*. Cambridge University Press, Cambridge, 1987. URL https://openlibrary.org/works/OL4962782W/Plans_and_Situated_Actions.

[67] Terry Winograd and Fernando Flores. *Understanding Computers and Cognition: A New Foundation for Design*. Ablex, Norwood, NJ, 1986.

[68] Edwin Hutchins. *Cognition in the Wild*. MIT Press, Cambridge, MA, 1995. URL https://mitpress.mit.edu/9780262082310/cognition-in-the-wild/.

[69] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–13. ACM, 2019. doi: 10.1145/3290605.3300233. URL https://doi.org/10.1145/3290605.3300233.

[70] Min Hun Lee and Chong Jun Chew. Understanding the Effect of Counterfactual Explanations on Trust and Reliance on AI for Human-AI Collaborative Clinical Decision Making. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2):1–22, 2023. doi: 10.1145/3610218. URL https://doi.org/10.1145/3610218.

[71] Anna Bodonhelyi, Efe Bozkir, Shuo Yang, Enkelejda Kasneci, and Gjergji Kasneci. User Intent Recognition and Satisfaction with Large Language Models: A User Study with ChatGPT. *arXiv preprint arXiv:2402.02136*, 2024. URL https://arxiv.org/abs/2402.02136.

[72] Kevin R. McKee. Human Participants in AI Research: Ethics and Transparency in Practice. *IEEE Transactions on Technology and Society*, 5(3):279–288, 2024. doi: 10.1109/tts.2024.3446183. URL https://doi.org/10.1109/tts.2024.3446183.

[73] Stefanos Nikolaidis, Yu Xiang Zhu, David Hsu, and Siddhartha Srinivasa. Human-Robot Mutual Adaptation in Shared Autonomy. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, pages 294–302. ACM, 2017. doi: 10.1145/2909824.3020252. URL https://doi.org/10.1145/2909824.3020252.

[74] Erin K. Chiou and John D. Lee. Trusting Automation: Designing for Responsivity and Resilience. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 65 (1):137–165, 2023. doi: 10.1177/00187208211009995. URL https://doi.org/10.1177/00187208211009995.

[75] Marcel Binz, Elif Akata, Matthias Bethge, Franziska Brändle, Fred Callaway, Julian Coda-Forno, Peter Dayan, Can Demircan, Maria K. Eckstein, Noémi Éltető, Thomas L. Griffiths, Susanne Haridi, Akshay K. Jagadish, Li Ji-An, Alexander Kipnis, Sreejan Kumar, Tobias Ludwig, Marvin Mathony, Marcelo Mattar, Alireza Modirshanechi, Surabhi S. Nath, Joshua C. Peterson, Milena Rmus, Evan M. Russek, Tankred Saanum, Johannes A. Schubert, Luca M. Schulze Buschoff, Nishad Singhi, Xin Sui, Mirko Thalmann, Fabian J. Theis, Vuong Truong, Vishaal Udandarao, Konstantinos Voudouris, Robert Wilson, Kristin Witte, Shuchen Wu, Dirk U. Wulff, Huadong Xiong, and Eric Schulz. A foundation model to predict and capture human cognition. *Nature*, 644(8078):1002–1009, July 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09215-4. URL http://dx.doi.org/10.1038/s41586-025-09215-4.

[76] Shirley Wu, Evelyn Choi, Arpandeep Khatua, Zhanghan Wang, Joy He-Yueya, Tharindu Cyril Weerasooriya, Wei Wei, Diyi Yang, Jure Leskovec, and James Zou. HumanLM: Simulating Users with State Alignment Beats Response Imitation, 2026. URL https://arxiv.org/abs/2603.03303.

[77] Tiancheng Hu, Joachim Baumann, Lorenzo Lupo, Nigel Collier, Dirk Hovy, and Paul Röttger. SimBench: Benchmarking the Ability of Large Language Models to Simulate Human Behaviors, 2026. URL https://arxiv.org/abs/2510.17516.

[78] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. URL https://arxiv.org/abs/2005.14165.

[79] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763, 2021. URL https://arxiv.org/abs/2103.00020.

[80] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. OpenVLA: An Open-Source Vision-Language-Action Model, 2024. URL https://arxiv.org/abs/2406.09246.

[81] Xinyi Xu, Zhaoxuan Wu, Rui Qiao, Arun Verma, Yao Shu, Jingtan Wang, Xinyuan Niu, Zhenfeng He, Jiangwei Chen, Zijian Zhou, Gregory Kang Ruey Lau, Hieu Dao, Lucas Agussurja, Rachael Hwee Ling Sim, Xiaoqiang Lin, Wenyang Hu, Zhongxiang Dai, Pang Wei Koh, and Bryan Kian Hsiang Low. Data-Centric AI in the Age of Large Language Models, 2024. URL https://arxiv.org/abs/2406.14473.

[82] Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Harsh Jha, Sachin Kumar, Lucy Li, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Abhilasha Ravichander, Dustin Schwenk, Oyvind Tafjord, Pete Walsh, Luke Zettlemoyer, Hannaneh Hajishirzi, Dirk Groeneveld, Noah A. Smith, Jesse Dodge, and Kyle Lo. Dolma: An Open Corpus of Three Trillion Tokens for Language Model Pretraining Research. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 15725–15788, 2024. URL https://arxiv.org/abs/2402.00159.

[83] Jeffrey Li, Alex Fang, George Smyrnis, Maor Ivgi, Matt Jordan, Samir Yitzhak Gadre, Hritik Bansal, Etash Guha, Sedrick Keh, Karan Arora, Shuran Zhao, Surya Jain, Sara Beery, Sebastien Bubeck, Alexandros G. Dimakis, Yair Carmon, Achal Dave, Ludwig Schmidt, and Vaishaal Shankar. DataComp-LM: In Search of the Next Generation of Training Sets for Language Models. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://arxiv.org/abs/2406.11794.

[84] Zhouhang Xie, Junda Wu, Yiran Shen, Yu Xia, Xintong Li, Aaron Chang, Ryan Rossi, Sachin Kumar, Bodhisattwa Prasad Majumder, Jingbo Shang, Prithviraj Ammanabrolu, and Julian McAuley. A Survey on Personalized and Pluralistic Preference Alignment in Large Language Models, 2025. URL https://arxiv.org/abs/2504.07070.

[85] Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. LaMP: When Large Language Models Meet Personalization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 7370–7392, 2024. doi: 10.18653/v1/2024.acl-long.399. URL https://aclanthology.org/2024.acl-long.399/.

[86] Ha Le, Rithika Lakshminarayanan, Jixin Li, Varun Mishra, and Stephen Intille. Collecting Self-reported Physical Activity and Posture Data Using Audio-based Ecological Momentary Assessment. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8 (3):1–35, August 2024. doi: 10.1145/3678584. URL https://doi.org/10.1145/3678584.

[87] Kaitlyn Zhou, Jena D. Hwang, Xiang Ren, Nouha Dziri, Dan Jurafsky, and Maarten Sap. REL-A.I.: An Interaction-Centered Approach To Measuring Human-LM Reliance. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11148–11167. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.naacl-long.556. URL https://aclanthology.org/2025.naacl-long.556/.

[88] Dionis Barcari, David Gamez, and Aliya Grig. Recording First-person Experiences to Build a New Type of Foundation Model, 2024. URL https://arxiv.org/abs/2408.02680.

[89] Martin Wolfgang Lauer-Schmaltz, Philip Cash, John Paulin Hansen, and Anja Maier. Towards the Human Digital Twin: Definition and Design – A survey, 2024. URL https://arxiv.org/abs/2402.07922.

[90] Chuhao Zhou and Jianfei Yang. HoloLLM: Multisensory Foundation Model for Language-Grounded Human Sensing and Reasoning, 2026. URL https://arxiv.org/abs/2505.17645.

[91] Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh Jing Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, Yitao Liu, Yiheng Xu, Shuyan Zhou, Silvio Savarese, Caiming Xiong, Victor Zhong, and Tao Yu. OSWorld: Benchmarking Multimodal Agents for Open-Ended Tasks in Real Computer Environments, 2024. URL https://arxiv.org/abs/2404.07972.

[92] Mariya Davydova, Daniel Jeffries, Patrick Barker, Arturo Marquez Flores, and Sinead Ryan. OSUniverse: Benchmark for Multimodal GUI-navigation AI Agents, 2025. URL https://arxiv.org/abs/2505.03570.

[93] Yucheng Li, Yunhao Guo, Frank Guerin, and Chenghua Lin. An Open-Source Data Contamination Report for Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 528–541. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.30. URL https://aclanthology.org/2024.findings-emnlp.30/.

[94] Cheng Xu, Shuhao Guan, Derek Greene, and M-Tahar Kechadi. Benchmark Data Contamination of Large Language Models: A Survey, 2024. URL https://arxiv.org/abs/2406.04244.

[95] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating Noise to Sensitivity in Private Data Analysis. In *Theory of Cryptography*, pages 265–284. Springer, 2006. doi: 10.1007/11681878_14. URL https://dblp.org/rec/conf/tcc/DworkMNS06.html.

[96] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pages 1273–1282. PMLR, 2017. URL https://proceedings.mlr.press/v54/mcmahan17a.html.