MIAI:

Human Misidentification of Artificial Agents in Computer-Mediated Communication

Jaeho Kim[1], Junehyuk Lee[2], and Christopher A. Baldassano[3]

[1]Program of Cognitive Science, Columbia University

[2]Department of Computer Science, Columbia University

[3]Department of Psychology, Columbia University

Abstract

As AI grows more human-like, how people judge its type-identity—human or not has become a critical question. Yet research remains fragmented across disciplines. This study addresses that gap by piloting an integrative framework that models type-identity judgment as an information-processing task. In a behavioral experiment, 44 participants held text-based conversations with AI agents varying in visual resolution, personality, and task framing. Participants answered recall questions, rated humanness, and reported judgment timing. Findings revealed that more content information does not lead to higher judgment accuracy. Less naturalistic avatars increased perceived humanness, especially when task awareness was low, and information recall declined when the type-identity judgment task was made explicit. A metacognitive mismatch was revealed: participants cited conversational content as central, yet neither content nor context models clearly outperformed the other. These results suggest type-identity judgments do not rely on conscious reasoning much more than subconscious heuristics. The study advances an integrative model of type-identity judgment, demonstrating how insights from prior literature may be theoretically and empirically revised and revisited.

Keywords: type-identity judgment, human–AI interaction, perceived humanness, computer-mediated communication, social cognition

Contents

MIAI:

Human Misidentification of Artificial Agents in Computer-Mediated Communication

A significant amount of research has been allocated to addressing the social implications of Artificial Intelligence (AI), for which the primary object of study has been AI as a general-purpose technology (Howard, 2019). Correspondingly, the major forerunners of AI technologies– such as OpenAI, Meta, and Google–have targeted developing artificial general intelligence (AGI) at an accelerated pace. At the same time, the commercialization of these technologies have also kept to speed, prompting non-tech oriented companies to attempt adopting and implementing them within their operations. Although they have not yet achieved the prophesied advent of AGI, they have made significant progress in developing and exposing to the public the capabilities of modern General Artificial Intelligence (GAI) and Large Language Models (LLMs).

Simultaneously, the essential generality of modern AI technologies—as well as the now-prevalent practice of lending out AI models via APIs—enable their adaptation to industrial sectors that are not necessarily centered around developing novel AI technology. Consequently, implementation and understanding have diverged, and the rift between the two are expected to widen if the current rate of progression in AI development is maintained (Al-Qudah, 2022). This has prompted genuine concerns over the impact of AI in terms of sustainability, stemming from the fundamental differences that exist between the human-developed systems that are accented by AI implementations and the natural systems from which a majority of conventional industries are derived (Khakurel et al., 2018).

Notwithstanding, the slogan of leading developers has remained the same: *Upgrade* and *Upscale*. The current practice of lending out AI technologies incentivizes the solving of complex problems with superior performance and robustness, which necessitates the development of increasingly complex and larger size AI models (Chen et al., 2022; Khakurel et al., 2018), without being limited by the divergence between implementation and understanding.

However, the criticism of such trends is *not* within the scope of the paper's aims. A large body of critical work already exists about the modern trend of divergence between technological implementation and scientific understanding, since long before the advent of contemporary AI (Heidegger, 1977; Kittler et al., 2018; Latour, 1993; Lyotard et al., 2010; McLuhan, 2002; Stiegler, 1998; Virilio, 2005; Winner, 1994). Instead, the project aims to direct attention towards a relatively underexplored topic in this matter: the advent of *human-like* AI. The increasing complexity and size of AI models tend to direct attention towards the prospects of AI that outperforms humans. That is, the focus of the debate around the implications of AI, due to its form of development, is inclined to be *better-than-human* AI. Within this focus, AI is conceptualized through a transhumanist lens. The question "What would the world look like when AI is omnipotent and omnipresent?" can be diluted into the question "What would the world look like when there are humans that are better than us?"

The current focus on the generality of AI technologies, paired with their exponential increase in the ability to *exceed* humans, results in another overlooked faculty: their ability to *replicate* humans. The ability for AI to be better-than-human likely implies their ability to be better human-like, in virtue of their ability to be better at performing any task, including the replication of human *behaviors*. This advent of human-like AI is not a projection. Explorations into the replication of human behaviors in AI implementation is already a present reality. This replication of human behaviors has been demonstrated to also affect human *judgment*, particularly in the realms of AI generated content (AIGC), which researchers in public policy have been endeavoring to address (Tao et al., 2023; Wittenberg et al., 2024; Zhang & Gosline, 2023). These replications have raised further ethical concerns, particularly in areas where AI is integrated into human-centered services, such as mental health care (Balcombe & De Leo, 2022; D'Alfonso, 2020; Denecke et al., 2021; Pham et al., 2022) and customer service (Fernandes & Oliveira, 2021; Hadi, 2019; Nicolescu & Tudorache, 2022).

It is imperative to recognize that being human-like is also a quality that revolves around human judgment, as the assessment of whether an AI is genuinely human-like essentially depends on the subjective, qualitative (and often social) evaluation of *bona fide* humans. It is also this non-quantifiable, qualitative property of human-likeness that differentiates the matter of human-like AI and better-than-human AI. The metrics and techniques that are developed to address the implications of better-than-human AI are often centered around factors that are outside the scope of individual human judgment. They are essentially macroscopic, materialist, and within objectivist dimensions, and the non-materialist, non-objectivist areas of evaluation—particularly ethics—have likewise stemmed their analyses from those dimensions (Alias Yaqoob & Robbins, 2024; LaCroix & Luccioni, 2025; Palumbo et al., 2024).

However, the metrics and techniques that will need to be developed for *human-like* AI require an essentially different set of dimensions. Most notably, whereas the metrics for *better-than-human* AI rely primarily on objectively measurable factors (such as the speed of computations or size of data processed) the metrics for *human-like* AI rely on subjectively felt, individual evaluations that are either reported or measured in human responses (such as perceptions of agency or Turing Test–style evaluations). Much work on this metrical distinction exists in the space of computational creativity (Agudo et al., 2022; Bown, 2021; Raj et al., 2023), and novel directions in experiment paradigms and apparatuses are being explored by academic and corporate entities (Floridi et al., 2009; Jannai et al., 2023; Meta Fundamental AI Research Diplomacy Team (FAIR) et al., 2022; Park et al., 2023; Warwick & Shah, 2016).

This turn toward mistaken identity highlights the emerging demand for a new cognitive ability: the need to perform *type-identity* judgment—the ability to determine *what* type of entity one is interacting with. Historically, humans have primarily relied on and frequently employed *token-identity* judgment: the capacity to infer *who* someone is, assuming that “someone,” in virtue of being able to hold a conversation at all, is by default human (Mahowald et al., 2023). Type-identity judgment, by contrast, involves determining whether an interlocutor is human at all, and it is this faculty that the advent of language-based AI now newly demands of us.

Yet, academic work on type-identity judgment remains fragmented and often secondary to other research agendas. Most prior research treats type-identity judgments as an explanatory or predictor variable for downstream effects like trust, affect, or interaction quality, rather than as a phenomenon worthy of investigation in its own right. Moreover, theoretical insights about this capacity are scattered across disciplines. Media studies frame it in terms of media affordances and interaction effects (Hancock et al., 2020; Littlejohn et al., 2017; Mieczkowski et al., 2021; Sundar, 2020; Walther, 1996; Weizenbaum, 1966); social cognition focuses on theory of mind and affect (Doyle et al., 2021; Gero et al., 2020; Peretti et al., 2023; Rato et al., 2022; Shank et al., 2019; Shneiderman & Muller, 2023; Tomasello, 2019; Q. Wang et al., 2021, 2024; Yadollahi et al., 2022); linguists emphasize pragmatic inference and conversational norms (Brachman et al., 2023; Loconte et al., 2025; Mahowald et al., 2023; Mieczkowski et al., 2021, 2021).

Altogether, what is missing is an integrative framework that consolidates these scattered theoretical intuitions into a single empirical target. Our study aims to offer such a framework by modeling type-identity judgment as a structured information-processing task and piloting it through a novel, modified Turing Test-style paradigm. Our larger project aims to propose a way by which we may be able to gain more significant traction, placing type-identity judgment at the center of theoretical and empirical investigation.

Theoretical Framework

To begin, this study adopts an *information-processing framework* to examine how individuals make type-identity judgments in virtual interactions. At the core of this framework is the distinction between content information and context information, grounded in how participants process cues during communication. The *interaction of interest* refers to the communicative exchange between agents—specifically, the dialogue that conveys information directly related to the identity of the interlocutors. For example, in a computer-mediated setting, this might include speech patterns, emotional engagement, or range of knowledge that is demonstrated by a chatbot during conversation.

Content information consists of signals embedded within the interaction of interest. It includes semantic, narrative, or linguistic details that can be logically mapped onto the true identity of the agent. *Context information,* in contrast, originates from outside the interaction of interest. This might include visual representations of the agent (e.g., their avatar or profile picture), interface layout, or participants' expectations and predispositions—factors that do not necessarily map on to the agent's identity. By isolating and analyzing the roles of content and context, this study aims to pilot a novel integrative framework that treats type-identity judgment as a structured, cue-based information-processing task. Through this lens, the study explores not only what types of information participants use, but how these cues are weighted, interpreted, and integrated—consciously or otherwise—during judgments of humanness.

Accordingly, the study employs a randomized within-subject design in which participants engage in structured conversations with AI chatbots in a simulated online interface.

Three key variables are experimentally manipulated:

(1) Personality (content): variations in chatbot knowledge access and linguistic behavior.

(2) Visual fidelity (context): the resolution of avatar images, modulated by pixelation.

(3) Task awareness (context): whether participants are aware that they will be making type-identity judgments.

Three quantitative dependent measures are collected:

(1) Rating: a numeric judgment of humanness (1–10).

(2) Judgment time: when in the conversation participants report having made their type-identity decision relative to the conversation duration.

(3) Sourcing: how much information participants recall and attribute their conversations with each interlocutor.

In addition, one qualitative measure is collected:

(1) Open-ended reasons for humanness ratings, later coded thematically to assess self-reported strategies and perceived cues.

Based on this framework and insights drawn from prior literature across media studies, social cognition, linguistics, we tested the following hypotheses:

(1) The availability of more information that is relevant to the true type-identity of interlocutors allows for more accurate type-identity judgments.

   ⇒ Therefore, slower judgments that have more content information available at hand will result in higher levels of accuracy.

(2) Naturalistic cues bias type-identity judgments towards ascriptions of human-like mental features.

⇒ Therefore, participants will rate interlocutors as more likely human when their visual representations are of higher resolution.

(3) Type-identity judgments are largely conscious processes of deliberation.

⇒ Therefore, participants' preferred type of information in their reasoning will correspond to the type that best predicts their judgments.

(4) Humans who are directly and actively involved in the creation of the referenced material are likely considered more knowledgeable and credible sources of information than non-human interlocutors who rely on second-hand acquired information in the form of documentation.

⇒ Therefore, participants will source more information from interlocutors they deem to be more likely human.

In addressing these questions, the study fills a gap between technical innovation and psychological understanding, offering a framework to assess and anticipate the social cognition risks posed by increasingly human-like AI.

Methods

Subjects

The study recruited 44 participants from the Columbia University psychology department subject pool (SONA). Participants were primarily undergraduate students, ensuring a consistent demographic for the exploratory pilot study. The study was conducted following IRB approval and adhering to the APA's *Ethical Principles of Psychologists and Code of Conduct*.

Apparatus

The study took place in a private, quiet room to minimize distractions and ensure consistent experimental conditions. Participants used a computer running a custom-built experimental interface created in Godot 4.3 using GDScript. Participants engaged in conversations with AI interlocutors whose responses are generated by the Llama 3.3 70B Instruct model (Grattafiori et al., 2024). These interlocutors were represented by images sourced from the Chicago Face Database (Ma et al., 2015), shown in three formats: full-resolution (high resolution), pixelated with 50px cell sizes (medium resolution), and pixelated with 100px cell sizes (low resolution). Pre-scripted questions about the conversationally acquired information and subjective humanness ratings were administered via the experimental interface.

The AI interlocutors were designed to simulate distinct personality types, with distinct settings across response generation parameters—temperature, top k, frequency penalty, and presence penalty—and prompts following a structured format: *Name, Affiliation, and Role within Game Development Team* → *Personality Traits* → *Conversational Style*. The prompts guided generation via the Llama 3.3 70B Instruct model, which instructed the LLM to impersonate each interlocutor according to how they would speak given their personality traits and conversational style, as well as what they would know given their role and affiliation as members of video game development team who are creating a game titled *Nostal;GEAR*. The AI interlocutors were made knowledgeable of the fiction video game through the use of a Retrieval-Augment Generation (RAG) pipeline, which provided the LLM with an information packet detailing the video game's game design, story, goals, and development status.

Abridged descriptions of each personality's system prompt are provided in Table 1; full prompts were longer and included additional behavioral cues. The full payload for each personality—including name, system prompt, and response generation parameters—are included in Appendix A. The information packet used for this RAG pipeline is included in Appendix B.

Table 1

Abridged Descriptions of Interlocutor Personalities

ID	Name	Role & Affiliation	Personality Traits	Conversational Style
0	Emily	Lead Game Designer, Columbia University	Analytical, detail-oriented, calm	Full sentences, proper punctuation, professional
1	David	Narrative Designer, New York University	Imaginative, passionate	Proper punctuation, polite
2	Samantha	Lead Programmer, Columbia University	Pragmatic, straightforward	Concise, skips unnecessary words, clear
3	Marcus	2D Artist, New York University	Creative, laid-back	Casual, enjoys brainstorming
4	Priya	Character Designer, Columbia University	Empathetic, artistic	Warm, conversational, uses encouraging language
5	Raj	Audio Engineer, New York University	Introspective, perfectionist	Thoughtful, occasionally clarifies or refines points
6	Jessica	Level Designer, Columbia University	Competitive, detail-driven	Confident, direct
7	Carlos	QA Lead, Columbia University	Practical, diligent	Pragmatic, clear communication
8	Ashley	Marketing Manager, New York University	Outgoing, charismatic	Energetic, uses exclamation points
9	Ryan	Junior Developer, New York University	Curious, enthusiastic	Informal

Note. These summaries are abridged versions of longer system prompts used to generate AI responses. Each prompt followed a consistent structural format across agents.

Procedure

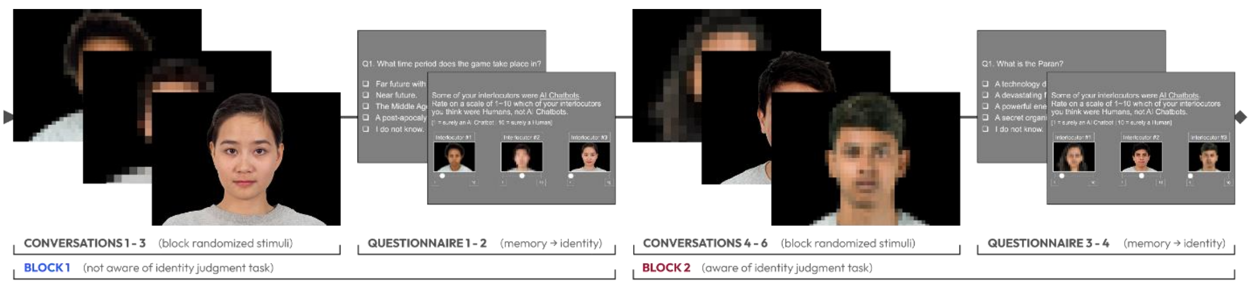
The design of this experiment was informed by two guiding principles (Figure 1):

(1) AI chatbots hold conversations with humans in an emulated, naturalistic computer-mediated communication setting.

(2) The conversations convincingly instill the belief that the interlocutors are *bona fide* human beings.

Figure 1

Experimental Procedure and Block Structure



Note. Participants engaged in six conversations with AI interlocutors presented in randomized visual conditions (high, moderation, or low resolution), divided into two experimental blocks. In Block 1, participants were unaware of the forthcoming type-identity judgment task; in Block 2, were aware of the type-identity judgment task due to the preceding type-identity judgment questionnaire. After each block, participants completed a memory questionnaire followed by humanness ratings and open-ended justifications for each interlocutor. Visual and task framing manipulations were counterbalanced across sessions.

Before beginning the experiment, each participant was instructed that they will be holding conversations with 6 interlocutors who are members of a team currently developing a video game. They were told that their task is to interview each interlocutor about the story of the video game the interlocutors are developing, with the topic ranging from plot, setting, and character. Once they entered the experiment room, they were provided with a sheet of paper that contained a list of potential questions about the video game's story that they may be asked about. Participants were advised to use this sheet as a guide for their conversations. These guides allowed the control of variance over the conversation content, keeping the conversations within a consistent flow. The lists of potential questions used distributed during the experiment are included in Appendix C.

Once the experiment began, the participants engaged in 6 text-based conversations with AI interlocutors, each lasting 6 minutes. The conversations were divided into two experimental blocks, with three conversations presented in each block. One participant completed only three conversations due to early termination. Each AI interlocutor was paired with one of three visual conditions (high, moderate, or low resolution). Participants conversed with the AI interlocutors in randomized order. Each response generated by the interlocutors also had a simulated delay, proportional to the length of the response generated (300 characters per minute) and with an additional random delay (between 2-4 seconds) added. A simulated connection screen preceded each trial to mimic live, synchronous interactions with human beings.

After every three conversations, participants answered 5 multiple-choice questions about the video game's story, including its plot, setting, and characters. Participants then rated how human-like they perceived each interlocutor to be on a 1–10 scale and briefly explained their reasoning in 1-2 sentences. They were also asked to indicate the point during the conversation at which they recalled making their judgment about the interlocutor's identity. After answering all of the questions presented after the end of the first block, participants received a new sheet of potential questions to use for reference during the second block. The questions asked during the length of the experiment are included in Appendix D. The total experiment duration was approximately 60–90 minutes, including introduction to the study and debriefing after the experiment.

Data Analysis

Regression Analysis

To examine how visual and conversational factors influenced participants' type-identity judgment and recall behavior, we conducted a series of regression analyses.

Linear Models. Ordinary least squares (OLS) regression models were fit to predict three dependent variables: perceived humanness ratings (*rating*), timing of humanness judgments (*timing_jdg*), and information sourcing (*sourcing*). Independent variables included visual resolution (*resolution*), experimental block (*block*), conversation order across the session (*position_in_session*), and conversation order within each block (*position_in_block*).

Separate models were constructed to examine three types of relationships.

First, models tested the main effects of individual predictors on sourcing behavior:

- *sourcing ~ block*
- *sourcing ~ position_in_block*

Second, models examined associations between dependent measures themselves:

- *rating ~ timing_jdg*
- *sourcing ~ rating*
- *sourcing ~ timing_jdg*

Third, models included interaction terms between block and other key predictors: *resolution*, *timing_jdg*, and *sourcing*. The interaction model most central to our analysis was:

- *rating ~ block * resolution*

Each model was evaluated based on standard inferential statistics, including estimated coefficients, standard errors, t-statistics, and p-values, as well as model fit indices such as R^2.

Mixed-Effects Models. To account for repeated measures within participants, we fit linear mixed-effects models with random intercepts for *subject_id*. These models mirrored the structure of the linear models, incorporating the same fixed effects and interaction terms. One particularly noteworthy model examined the relationship between rating and judgment time:

- *timing_jdg ~ rating + (1 | subject_id)*

Exploratory analyses revealed notable between-subject variability in the timing of humanness judgments (*timing_jdg*), justifying the use of mixed-effects modeling for this outcome. In contrast, models predicting *rating* and *sourcing* exhibited singular fit, indicating that between-subject variance in those measures was negligible. Model fit was assessed using restricted maximum likelihood (REML) estimation, AIC, and BIC, and additional checks for singularity were conducted to evaluate model stability.

Sourcing

To assess recall sourcing, we developed an automated content-sourcing procedure that compared participants' multiple-choice responses to their prior conversation transcripts. Each participant engaged in conversations with six AI interlocutors, each responsible for presenting different components of a fictional story. After these interactions, participants answered two sets of five story comprehension questions which corresponded to content embedded in the transcripts from interlocutors 1–3 and 4–6, respectively.

For each comprehension question, the answer options were mapped onto canonical story facts (e.g., "The story takes place in a post-apocalyptic future"). Each selected response was then evaluated against the three relevant transcripts from the immediately preceding block using

OpenAI's o3-mini model, a lightweight instruction-tuned language model accessed via the OpenAI API. The model was prompted with the relevant transcript segment, the comprehension question, and the participant's chosen answer, and was instructed to determine whether the content of the answer was verifiably present in the transcript. The model responded with either "Yes" or "No."

This procedure was repeated for each relevant transcript per question—specifically, the three interlocutors who held conversations before the corresponding comprehension assessment. All automated evaluations were manually reviewed by the lead researcher to ensure alignment with the intended scoring logic. Manual verification confirmed the consistency and accuracy of the model's output: no false positives or false negatives were observed. In every instance, the model's binary judgment matched the lead researcher's own assessment of whether the transcript adequately supported the selected answer, indicating a high degree of reliability in the automated pipeline.

Each transcript was then assigned a "sourcing" count whenever it was identified as containing information that justified a selected answer. Consequently, each transcript received a discrete rate ranging from 0 to 5 (as there were 5 recall questions per block), representing the number of times it served as the source for a correct or supported answer. These sourcing rates were used as dependent variables in subsequent linear and mixed-effects regression models to quantify how often each AI interlocutor contributed to participants' understanding of the game's story. This automated sourcing approach provides a scalable and replicable method for assessing information recall, allowing for semantic flexibility in participant responses while grounding the scoring system in clear, verifiable alignments between information recall and transcript data.

Thematic Analysis

Open-ended responses in which participants explained their humanness judgments were analyzed through a combination of LLM-assisted code generation, manual thematic assignment, and statistical modeling. The initial step involved feeding the full set of participant explanations into an instruction-tuned LLM to identify potential recurring themes. Based on the LLM's output, the lead researcher manually reviewed and refined the coding schema, defining a set of thematic categories that captured the most salient interpretive patterns. These categories included traits associated with being human-like or AI-like, conversational style, level of engagement or effort, and expressions of uncertainty.

Each participant response was then manually assigned to one or more of these themes by the lead researcher. This allowed for multi-label annotation, recognizing that a single explanation could reference multiple types of reasoning. Prior to coding, all text responses were cleaned and tokenized to prepare for downstream frequency analysis and visualization. To facilitate group-level interpretation, the distribution of themes was examined across two axes: interaction order (i.e., the position of the conversation within the session) and humanness rating (binned into quantiles). These distributions were visualized using bar plots and word clouds to highlight common language and reasoning patterns.

In parallel with the thematic coding, a higher-level binary classification scheme was applied to each response. According to our theoretical framework, this scheme categorizes responses as content-based, context-based, or ambiguous (expressing uncertainty or non-informative reasoning). This binary coding enabled the analysis of broader interpretive strategies across participants. To test for differences in the use of interpretive strategies across experimental conditions, we modeled the binary presence of each strategy using logistic regression. The likelihood of using a particular strategy was predicted based on experimental block (Block 1 vs. Block 2), enabling us to evaluate how task awareness shaped participants' reasoning.

Model Comparison

To evaluate the relative contributions of conversational content and contextual visual information to humanness judgments, we constructed and tested two competing linear regression models. The first, referred to as the context model, incorporated experimental manipulations of avatar resolution, task awareness (block), and the interaction between these two factors. The second, the content model, included a categorical variable representing interlocutor meta-personality. These meta-personality clusters were generated by grouping chatbot profiles into four distinct categories based on shared linguistic traits and response generation behavior.

The groupings were determined through transcript inspection and lexical analysis. Initially comprising five clusters, the "Imaginative" and "Upbeat" categories were collapsed into a single Expressive group due to their shared conversational tone, high use of exclamation marks, and verbosity. This resulted in four meta-personality levels:

(1) *Measured:* Personality IDs 0 and 2

(2) *Expressive:* Personality IDs 1, 3, 8, and 9

(3) *Gentle:* Personality IDs 4 and 5

(4) *Direct:* Personality IDs 6 and 7

Examples of the transcripts produced by each meta-personality are shown in Table 2. Full transcripts are included in Appendix E (Table E2). This structure balances interpretability with statistical parsimony, maintaining 3 degrees of freedom to match the complexity of the context model.

Table 2

Transcript Examples from Each Interlocutor Meta-Personality

Meta-Personality	Example Personality ID	Example Transcript Excerpt
Efficient	2	User: What is the general setting of the game and how does it contribute to the plot? Interlocutor: Game's set in Aris City, a post-apocalyptic metropolis rebuilt after a disaster, and it explores themes of nostalgia, trauma, and human resilience. User: who is the character and what is their background story Interlocutor: Protagonist has fragmented memories, trying to recall past in a world that's been devastated by ecological disaster.
Expressive	8	User: hello would you like to tell me about your game Interlocutor: Oh my gosh, yes! I'd love to tell you about Nostal;GEAR, it's a 2D puzzle platformer with a really deep narrative! User: whats the plot Interlocutor: It's set in a post-apocalyptic world where a disaster called Paran swept across the globe, and players navigate Aris City to uncover the truth behind a series of murders!
Gentle	4	User: Hello! Interlocutor: Hey, how's it going? User: I'm doing well! I just have a few more questions about your game-- the setting (remind me of the name), what are its key features? Interlocutor: The setting is Aris City, and key features include puzzle-solving, platforming, and a narrative-driven story.

Meta-Personality	Example Personality ID	Example Transcript Excerpt
Direct	7	User: hello Interlocutor: hey, what's up User: would you like to tell me about the story of your game? Interlocutor: it's about a world that got devastated by a disaster called Paran, and the player tries to uncover the truth behind a series of murders in a city called Aris City.

We assessed model performance using several evaluation metrics. Adjusted R^2 was used to account for model complexity and to quantify the proportion of variance explained by each model. To further compare model fit while penalizing for the number of parameters, we computed two information-theoretic indices: the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). To assess the generalizability of the models beyond the training data, we also implemented five-fold cross-validation. For each fold, we calculated the root mean square error (RMSE) and adjusted R^2 to evaluate model accuracy and stability across different subsets of data.

In addition to overall comparisons, we conducted block-wise analyses to examine whether model performance varied between Block 1 (low task awareness) and Block 2 (high task awareness). For each block, we recalculated adjusted R^2 and BIC to evaluate how well each model captured variance under different cognitive framing conditions. These results were then visualized to track model performance across the two experimental blocks. Taken together, this modeling strategy allowed us to directly compare the predictive strength of content information cues and context information cues in shaping humanness judgments. It also enabled a more nuanced understanding of how task awareness may moderate the relative influence of these cues.

Results

Design Validation

The experimental apparatus functioned reliably across all sessions, having been deployed using a stable version of the custom-built software environment. Conversations between participants and AI interlocutors occurred in a controlled, naturalistic computer-mediated communication setting designed to emulate typical online interactions. To ensure experimental fidelity, each session was screen-recorded and later reviewed by researchers to verify that all interactions unfolded as intended. These reviews confirmed that the simulation consistently reproduced plausible, synchronous dialogues between participants and seemingly human partners.

Participant responses further supported the ecological validity of the setup. The average humanness rating was 45.66 (SD = 29.52), with a median of 49.00 and a mode of 50. Given the 0–100 rating scale, these values suggest that participants often perceived the AI interlocutors as plausibly human, with ratings clustering around the midpoint. In fact, 50% of judgments were equal to or above 50, reinforcing the interpretation that participants were frequently misled into believing their interlocutors were bona fide human beings. Moreover, the wide variability in responses implies a diversity of subjective impressions and evaluative criteria.

This diversity was also reflected in participants' written explanations for their judgments. These qualitative responses provided an additional layer of validation for the study's experimental framing, demonstrating that participants engaged meaningfully with the task and articulated their inferences based on both concrete and intuitive conversational cues. To illustrate how conversational features corresponded with participants' humanness ratings, Table 3 presents three example reasons alongside their assigned ratings and interlocutor personalities. The data rows including their interlocutor personality, meta-personality, rating, and participant justification are included in Appendix E (Table E1).

Many referenced specific conversational features or behaviors as cues of humanness or artificiality. Conversely, low-rated interlocutors were often flagged for lacking coherence or responsiveness. Uncertainty was also common when cues for either identity were noted by the participants. The examples highlight contrasts in conversational tone, elaboration, and engagement style, which participants appeared to use as cues in forming type-identity judgments.

Table 3

Examples of High and Low Humanness Ratings with Participant Justifications

Rating	Interlocutor ID (Meta-Personality)	Participant Justification
78 (High)	1 (*Expressive*)	"[T]his has to be human—the conversation and their responses actually made me more intere[s]ted in the game to the point where [I] wanted to ask more questions. The person probably played it already and knows the game good,"
43 (Moderate)	0 (*Efficient*)	"This felt like the most normal conversation of them all, but I could also see how it could be AI through the super descriptive answers. This, however, also just made me feel like I was talking to someone who put a lot of time and effort into something they are proud of."

Rating	Interlocutor ID (Meta-Personality)	Participant Justification
12 (Low)	6 (Direct)	“The answers given were a bit weird. Sometimes they didn't make sense or j[us]t didn't answer my question. If I asked a continuing question it didn't understand,”

Note. Each row includes the humanness rating, interlocutor ID, and participant justification. These examples reflect how participants interpreted linguistic cues—such as tone, coherence, and affective engagement—when judging whether an interlocutor was human.

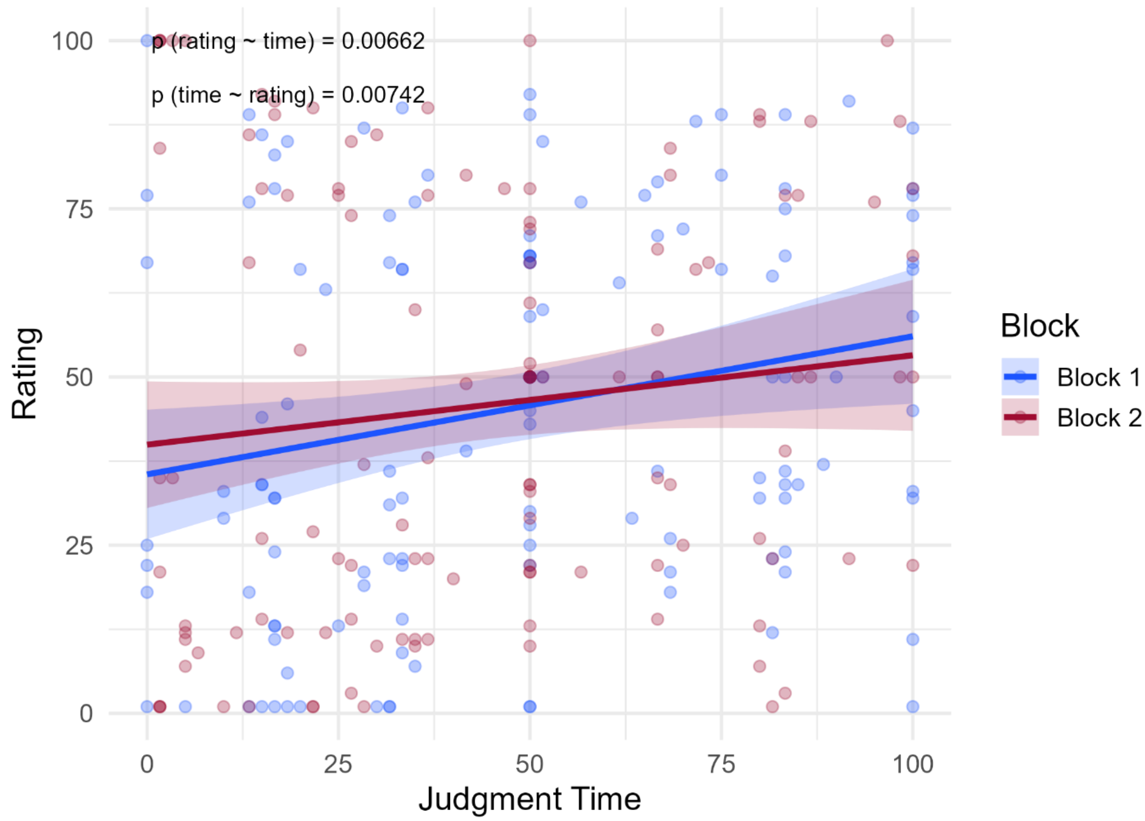
Finally, the experimental manipulation of task awareness proved effective. A significant decrease in information sourcing was observed in Block 2, when participants were explicitly informed of the type-identity judgment task. This decline, supported by post-experiment interviews, suggests a shift in attentional focus from information retention to identity evaluation. However, we acknowledge a potential confound: task awareness was introduced concurrently with increased time into the session, raising the possibility that participants had already exhausted their most salient recall questions by Block 2.

To mitigate this, participants were given a new list of potential questions before Block 2 to encourage fresh lines of inquiry. Still, sourcing declined sharply—and this was accompanied by a substantial rise in non-informative responses. Specifically, the rate of “*I don’t know*” responses increased from 4.14% in Block 1 to 14.3% in Block 2, and several selected answers did not match content present in any prior conversation. These patterns suggest that while task framing clearly influenced participants’ cognitive priorities, temporal and motivational factors may have contributed to the observed decline in information sourcing as well.

Content-Level Effects

Figure 2

Humanness Rating as a Function of Judgment Timing by Block



Note. Linear regression lines showing the relationship between judgment timing and humanness rating across experimental blocks. In both Block 1 and Block 2, longer deliberation times were associated with higher humanness ratings, though the effect was stronger in Block 1.

To examine how the availability of conversational content influenced judgments of humanness, we analyzed the relationship between participants' reported judgment timing and their humanness ratings. In a linear model predicting humanness rating from judgment time (*rating ~ timing_jdg*), judgment time was a significant positive predictor (β = 0.17, SE = 0.06, t = 2.74, p = .007). These results, visualized in Figure 2, indicate a positive correlation:

participants tended to give higher humanness ratings when they took longer to arrive at a decision, suggesting that more content information might not necessarily reduce the rate of misidentification. However, this finding contradicts Hypothesis 1, which predicted that slower judgments—interpreted as increased availability of content information—would yield more accurate and potentially lower ratings.

Further analysis tested whether the relationship between judgment timing and humanness rating varied across task awareness conditions using a linear regression model that included an interaction term between timing and block (*rating ~ timing_jdg * block*). The main effect of judgment timing on humanness rating was positive but not statistically significant ($\beta = 0.28$, $SE = 0.19$, $t = 1.43$, $p = .15$). The effect of block was similarly non-significant ($\beta = 4.41$, $SE = 6.83$, $t = 0.65$, $p = .52$). Crucially, the interaction between timing and block was also non-significant ($\beta = –0.07$, $SE = 0.13$, $t = –0.58$, $p = .56$), indicating that the relationship between judgment time and humanness rating did not differ meaningfully between low task awareness conditions (Block 1) and high task awareness conditions (Block 2). This suggests that content-level effects may be considerably resistant to changes in task awareness.

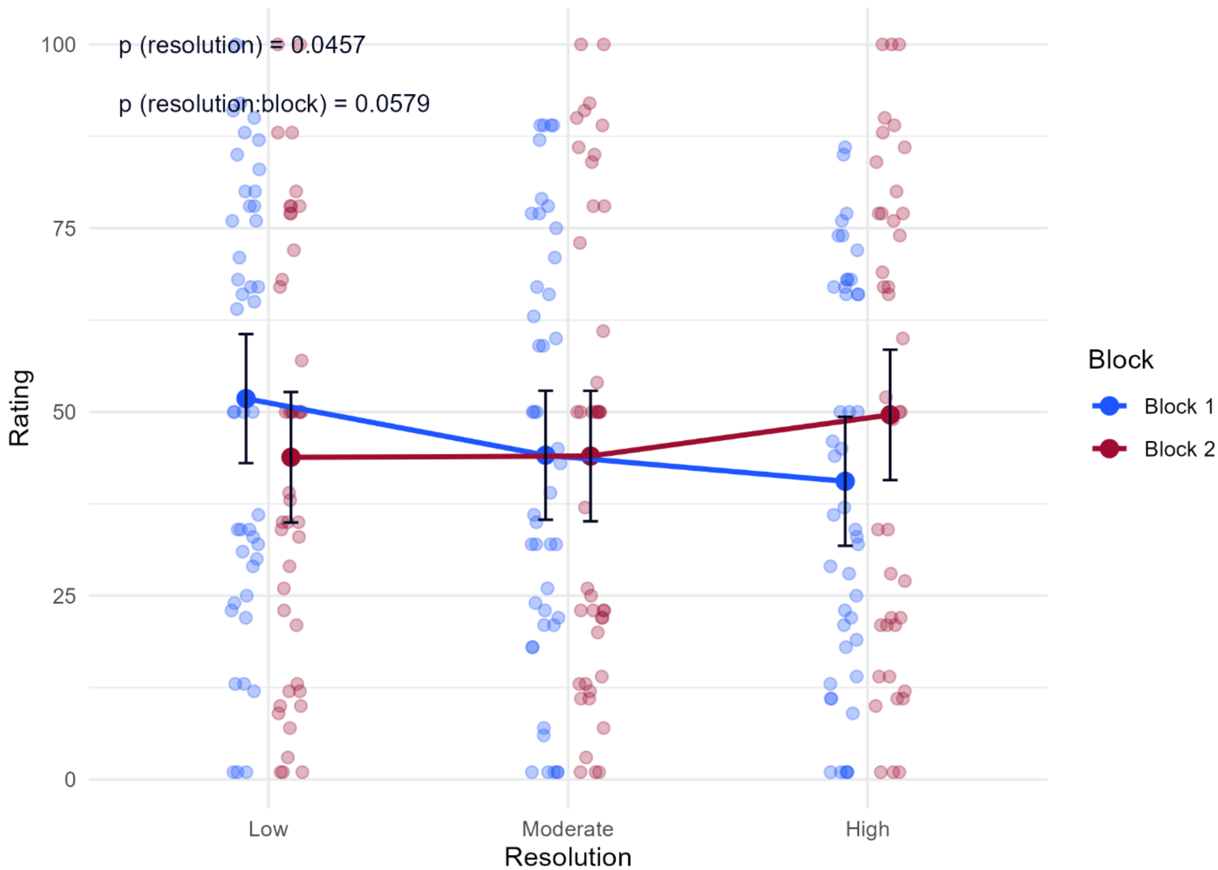
Together, these findings challenge the notion that increased content exposure necessarily improves judgment accuracy. Even though longer deliberation was associated with higher humanness ratings, this effect did not significantly differ across task awareness conditions. This suggests that extended deliberation may not reflect more accurate inference, but rather epistemic uncertainty or over-interpretation—leading participants to project humanness where it may not exist. These results underscore the complexity of content-based processing in type-identity judgment and suggest that the accumulation of interactional detail can sometimes inflate, rather than clarify, impressions of humanness.

Context-Level Effects

To investigate how contextual cues influence type-identity judgments, we analyzed the effects of manipulated visual resolution and participants' task awareness on humanness ratings. Visual resolution was operationalized across three levels: *high*, *moderate*, and *low*. Task awareness was manipulated across two blocks: unaware of type-identity judgment task in Block 1 and aware of the task in Block 2.

Figure 3

Effects of Visual Resolution on Humanness Ratings by Block



Note. Mean humanness ratings across three visual resolution conditions (Low, Moderate, and High), separated by block (Block 1 = unaware of type-identity judgment task; Block 2 = aware). Ratings were highest for low-resolution images in Block 2 and lowest under high-resolution images in Block 1.

Resolution

Participants rated AI interlocutors with lower-resolution, more pixelated visual representations as more likely to be human (Figure 3). In the linear regression model predicting humanness ratings from block and resolution (*rating ~ block * resolution*), resolution emerged as a significant predictor ($\beta = 14.13$, $SE = 7.04$, $p = .046$). Specifically, less naturalistic visual resolution—that is, more distorted, pixelated images—was associated with higher humanness ratings. This finding runs counter to Hypothesis 2, which had predicted that greater visual fidelity would promote humanness attribution by more closely aligning with real-world perceptual expectations.

One interpretation of this result is that pixelation may obscure subtle but revealing cues of artificiality that are more visible in high-resolution avatars. By reducing visual detail, pixelation may prevent participants from detecting heuristic signs that an interlocutor is not human—thus inadvertently increasing humanness attributions. Another possible explanation is that pixelation serves as a contextual cue linked to everyday digital communication practices, such as low-resolution webcam feeds. In such cases, participants may unconsciously associate pixelation with authentic, casual human presence.

Together, these findings challenge conventional assumptions that more realistic visual rendering necessarily enhances believability. Instead, degraded visual fidelity may paradoxically increase the likelihood of being perceived as human by either concealing imperfections or mimicking familiar, low-bandwidth human interaction environments.

Task Awareness

The effect of visual resolution on humanness ratings was moderated by participants' awareness of the type-identity judgment task. In Block 1—when participants were not informed that they would later be asked to evaluate their interlocutors' identities—the influence of resolution was pronounced. Conversely, in Block 2, where participants were explicitly instructed to judge humanness, the resolution effect diminished. This attenuation was reflected in a marginally significant interaction between resolution and block ($\beta = -8.51$, $SE = 4.47$, $p = .058$), suggesting that task awareness reduced reliance on heuristic visual cues.

Although the main effect of block was not statistically significant ($\beta = 8.82$, $SE = 5.77$, $p = .127$), the interaction pattern supports the interpretation that participants shifted from automatic, context-based heuristics to more effortful, content-based reasoning when identity evaluation became a salient task. Post-experiment interviews further corroborate this interpretation: several participants in Block 2 reported intentionally disregarding visual features—such as avatar realism or image quality—in favor of focusing on the substance of the conversation.

This shift indicates that task framing can significantly influence the type of cues participants attend to when making type-identity judgments, especially in computer-mediated settings where both context and content are available but separable. While content-level effects appeared more resistant to changes in task framing, context-level effects—such as those driven by visual resolution—were more sensitive to task awareness. This divergence suggests that type-identity judgments may rely on different cognitive routes depending on the nature of the cue, with contextual heuristics being more easily disrupted by explicit evaluative goals.

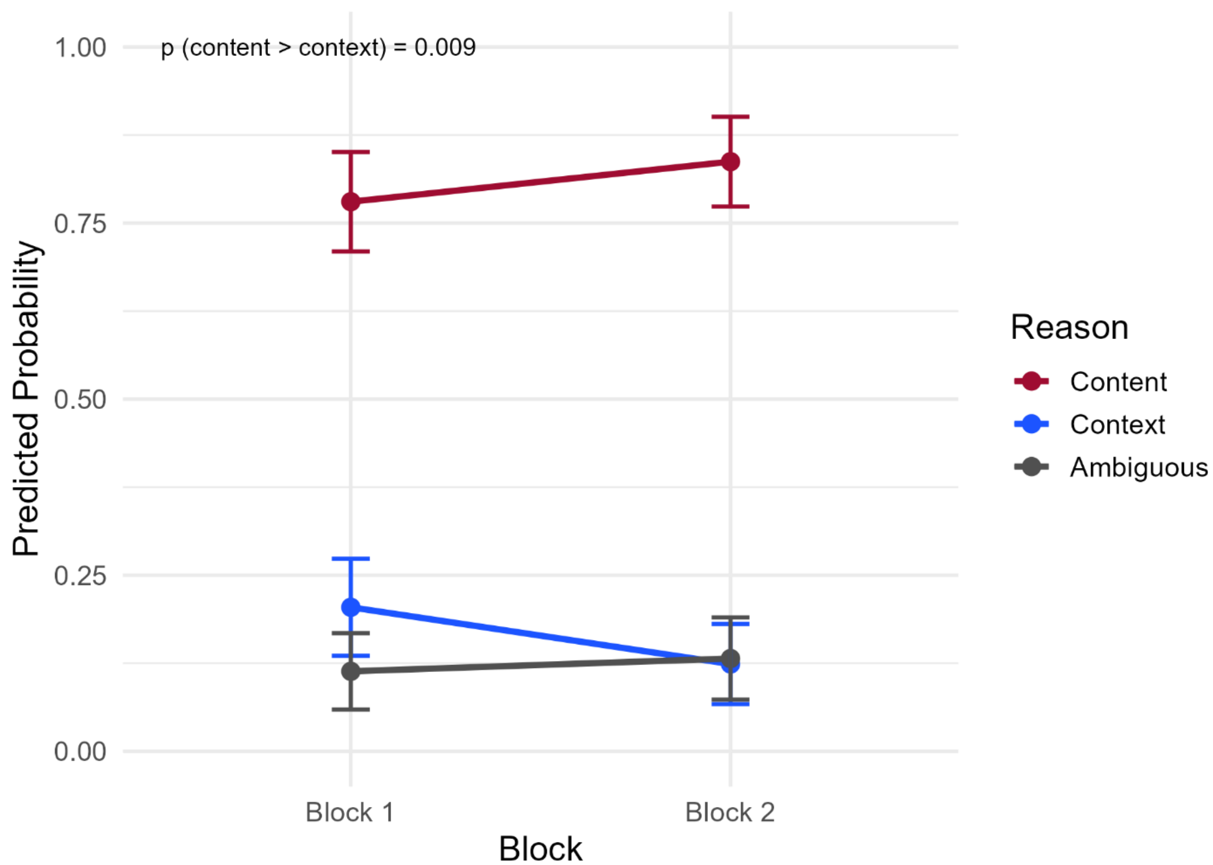
Metacognition

Self-Reporting Preference

To further examine participants' introspective reasoning, our primary approach was to thematically code their open-ended justifications for humanness ratings into three mutually non-exclusive interpretive categories: content-based, context-based, and ambiguous. Content-based responses referenced elements of the conversation itself—such as language fluency, coherence, or depth of information. Context-based explanations drew on peripheral or structural cues, such as avatar image resolution or response delay. Ambiguous responses expressed uncertainty or offered non-specific justifications (e.g., "just a feeling"). Because participants often referenced more than one type of cue, a single response could be tagged with multiple categories; as a result, proportions do not sum to 100%.

Figure 4

Self-Reported Interpretive Strategies by Block



Note. Proportion of participant explanations tagged as content-based, context-based, or ambiguous across Blocks 1 and 2. Content-based reasoning dominated overall and increased slightly when type-identity judgment was an explicit task (Block 2). In contrast, context-based justifications decreased.

As visualized in Figure 4, across all coded responses, content-based reasoning was by far the most prevalent, appearing in 80.9% of justifications. Context-based reasoning appeared in 16.4%, while 12.3% of responses were marked as ambiguous. Notably, this trend intensified under explicit task framing: in Block 2, where participants knew they would be making type-identity judgments, the use of content-based reasoning rose to 83.7%, while context-based reasoning fell to 12.4%. In Block 1, content and context-based tags occurred at 78.0% and 20.5%, respectively.

These findings suggest that participants not only favored content information but considered themselves to have increasingly relied on them when identity evaluation was made salient. This aligns with prior research indicating that individuals tend to place more trust in semantically rich, narrative-driven content when making social judgments—even in artificial or ambiguous settings. However, as discussed in the section below, this preference did not translate into stronger predictive power for content-based features in our computational models. While participants believed they were making content-driven evaluations, their ratings may have been influenced by factors beyond conscious access or reportability.

Model Alignment

To evaluate whether participants' humanness ratings were better predicted by content or context information, we compared two linear regression models. The content model (*rating ~ meta_personality*) included interlocutor meta-personality clusters, derived from shared linguistic traits and response generation behavior across interlocutor personalities. The context model (*rating ~ block * resolution*) included avatar resolution, task awareness, and their interaction.

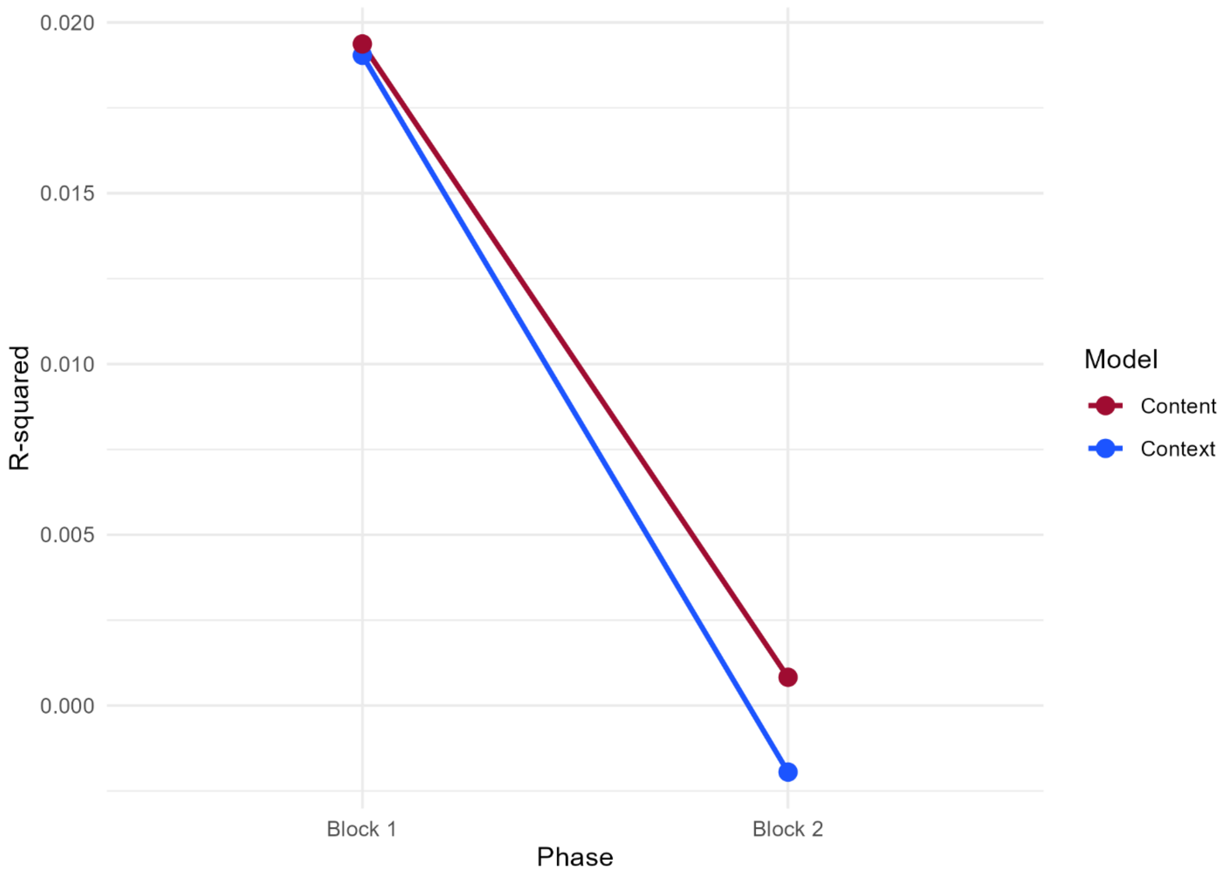
Importantly, the construction of these models was grounded in both participants' self-reported reasoning strategies and the primary variables manipulated in the study design. The factors included in each model correspond to the features that participants frequently cited when explaining their ratings. Meta-personality—and the original interlocutor personalities—influenced conversational tone, information quality, and phrase patterns, which were common themes in content-based reasoning, while visual resolution and timing of task awareness were often mentioned in context-based explanations. Although these models are not exhaustive representations of all possible influences on humanness ratings, they allow us to systematically compare the predictive value of the core experimental variables we directly manipulated.

Content-based reasoning was the most frequently cited strategy in open-ended justifications, particularly under high task awareness (Block 2), suggesting that participants believed they were engaging in deliberative, semantic analysis. Contextual cues, in contrast, were reported less frequently and typically associated with lower task salience. As such, we hypothesized that the content model would strongly outperform the context model in predicting humanness ratings—mirroring participants' strong preference for content information and aligning with the assumption that type-identity judgments is a conscious reasoning process.

However, neither model accounted for a substantial proportion of variance in humanness ratings (Figure 5). The context model yielded an adjusted R^2 of 0.004, with resolution emerging as the only statistically significant predictor ($\beta = 14.13$, SE = 7.04, $t = 2.008$, $p = .046$). The effects of block ($\beta = 8.82$, SE = 5.77, $t = 1.528$, $p = .127$) and the resolution × block interaction ($\beta = –8.51$, SE = 4.47, $t = –1.904$, $p = .058$) were not significant, although the interaction approached marginal significance. The content model performed only slightly better, with an adjusted R^2 of 0.009. However, none of the meta-personality categories significantly predicted humanness ratings. For example, the *Expressive* group had a coefficient of $\beta = 5.01$ ($p = .322$), *Gentle* $\beta = 2.13$ ($p = .712$), and *Direct* $\beta = –6.16$ ($p = .278$), offering minimal explanatory value.

Figure 5

Adjusted R^2 for Content and Context Models by Block

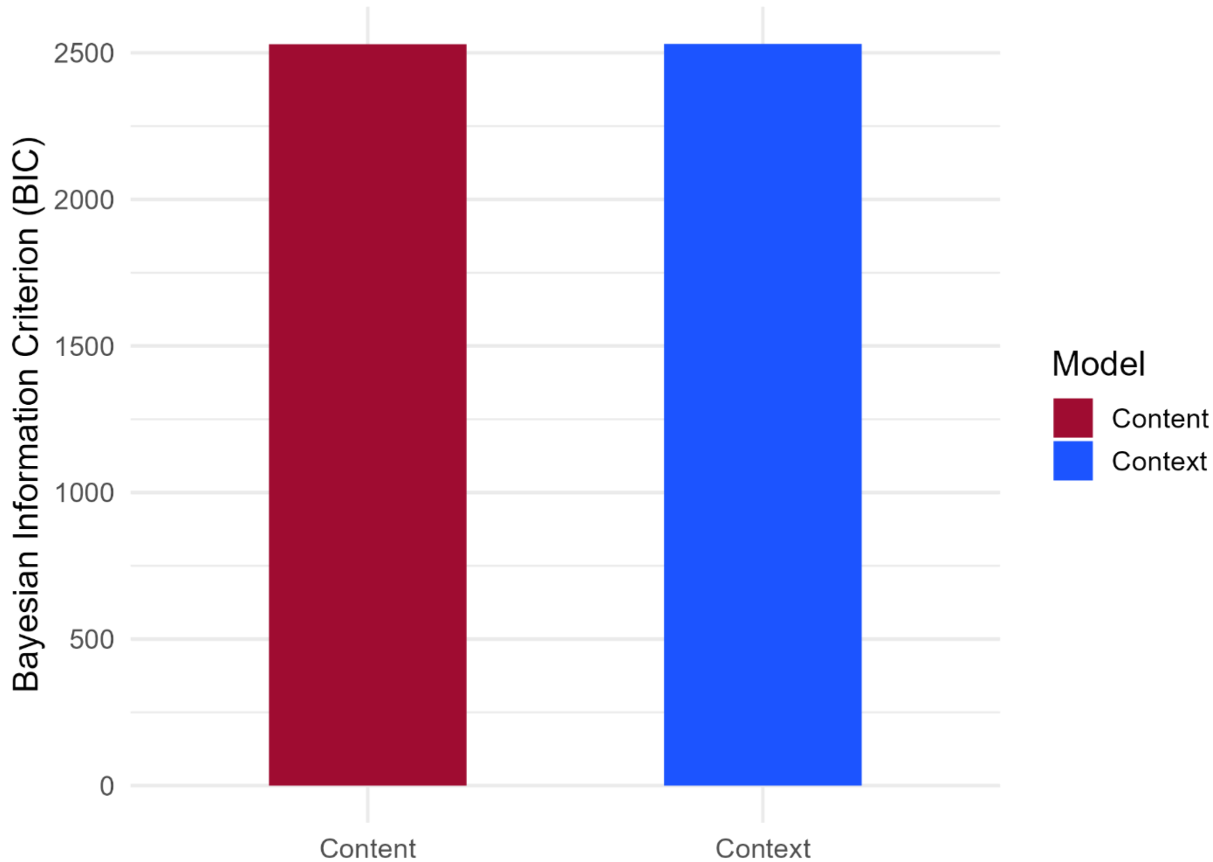


Note. Adjusted R^2 of content and context models across Blocks 1 and 2. Both models showed stronger predictive power in low task awareness conditions (Block 1) than in high task awareness conditions (Block 2).

We further examined explained variance by task awareness. Interestingly, both models performed better in Block 1—when participants were not aware that they would be making type-identity judgments—than in Block 2. This suggests that type-identity judgments made without explicit task awareness may rely more heavily on the specific cues we manipulated (e.g., resolution and personality style), whereas task awareness may trigger other strategies or cue integrations that were not well captured by our current models.

Figure 6

BIC Comparison Between Content and Context Models



Note. BIC comparisons between content and context models. The content model showed marginally better fit (lower BIC) than the context model, consistent with a Bayes Factor (BF ≈ 1.94) indicating weak evidential support.

Model fit comparison supported the overall pattern (Figure 6). The content model showed slightly better fit (AIC = 2511.30; BIC = 2529.13) than the context model (AIC = 2512.64; BIC = 2530.46), but the differences (ΔAIC = 1.34; ΔBIC = 1.34) were negligible and fall below thresholds typically considered meaningful for model selection. When converted to a Bayes Factor, this yielded a BF ≈ 1.94—indicating that the content model is only weakly more likely than the context model, offering minimal evidential support.

These findings contradict Hypothesis 3, which predicted alignment between self-reported reasoning and predictive model performance. Instead, the results point to a metacognitive mismatch: while participants overwhelmingly claimed to rely on content information cues—particularly when they were aware of the type-identity judgment task—this preference was not mirrored in the relative strength of the content model. This discrepancy suggests that participants' introspective access to their own identity reasoning processes may be limited, and that even seemingly deliberative, content-driven judgments may initially be shaped by unconscious heuristics or cue integration mechanisms not easily articulated or captured by our current, Turing Test–based experimental design.

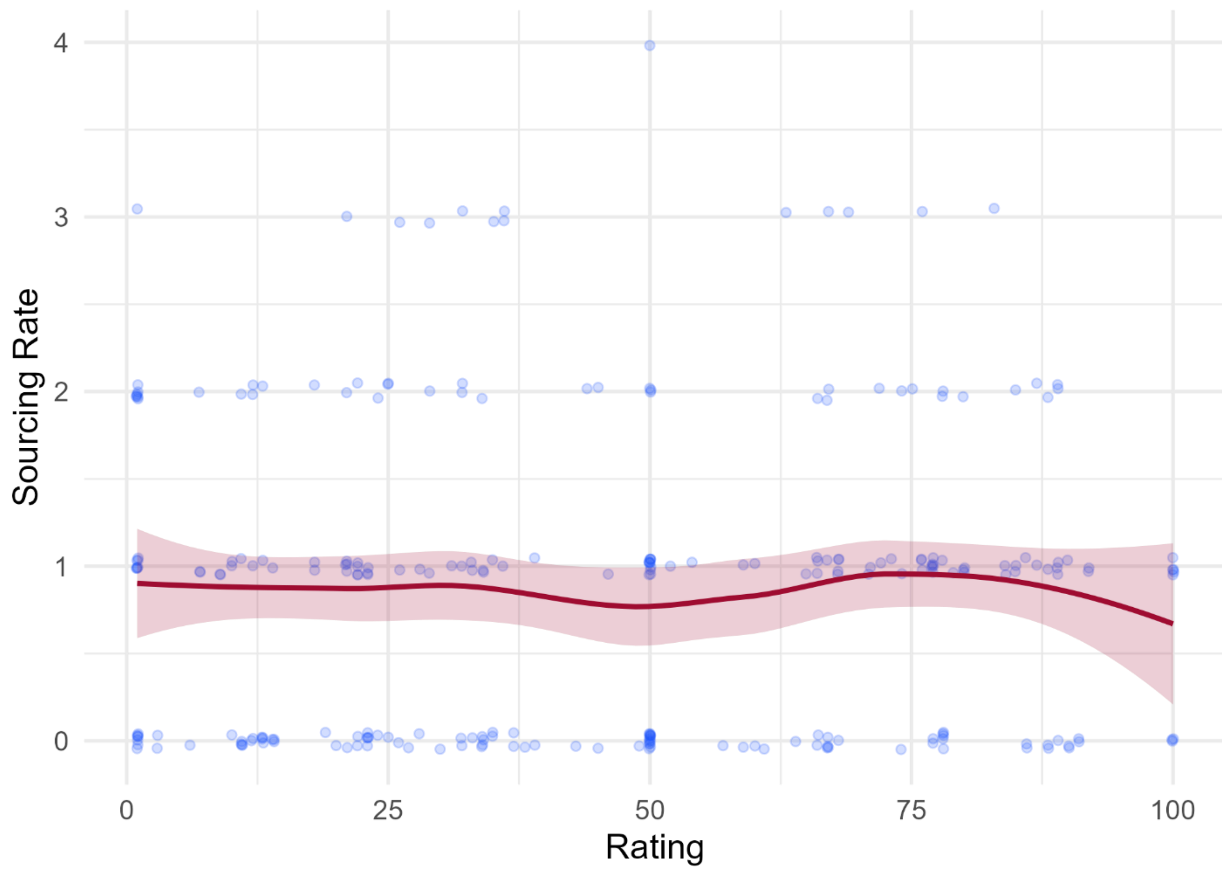
While these findings suggest no clear dominance between the two models, they also highlight the preliminary nature of this comparison. Both models are relatively coarse-grained, and further refinement—such as incorporating interactional dynamics, linguistic variability metrics, or richer context features—will be necessary to reach a clearer understanding of what factors drive type-identity judgments in computer-mediated conversations.

Sourcing

To assess whether perceived identity influenced participants' information-gathering behavior, we tested whether humanness ratings and judgment times predicted information sourcing—the extent to which participants recalled and attributed information to a given interlocutor. Contrary to Hypothesis 4, no significant associations were observed.

Figure 7

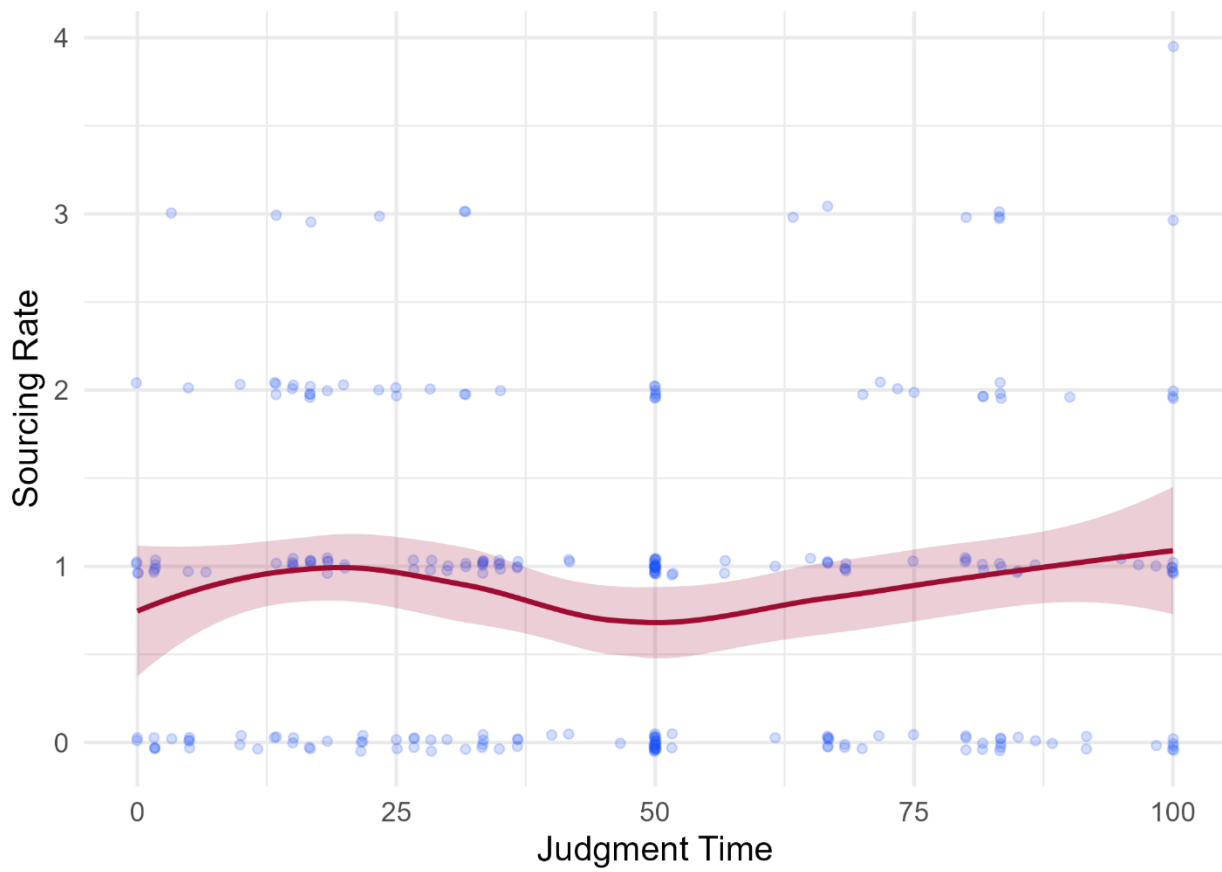
Sourcing by Humanness Rating



Note. Linear regression lines showing the relationship between rating and sourcing. No significant association was found between humanness ratings and sourcing behavior.

Figure 8

Sourcing by Judgment Time

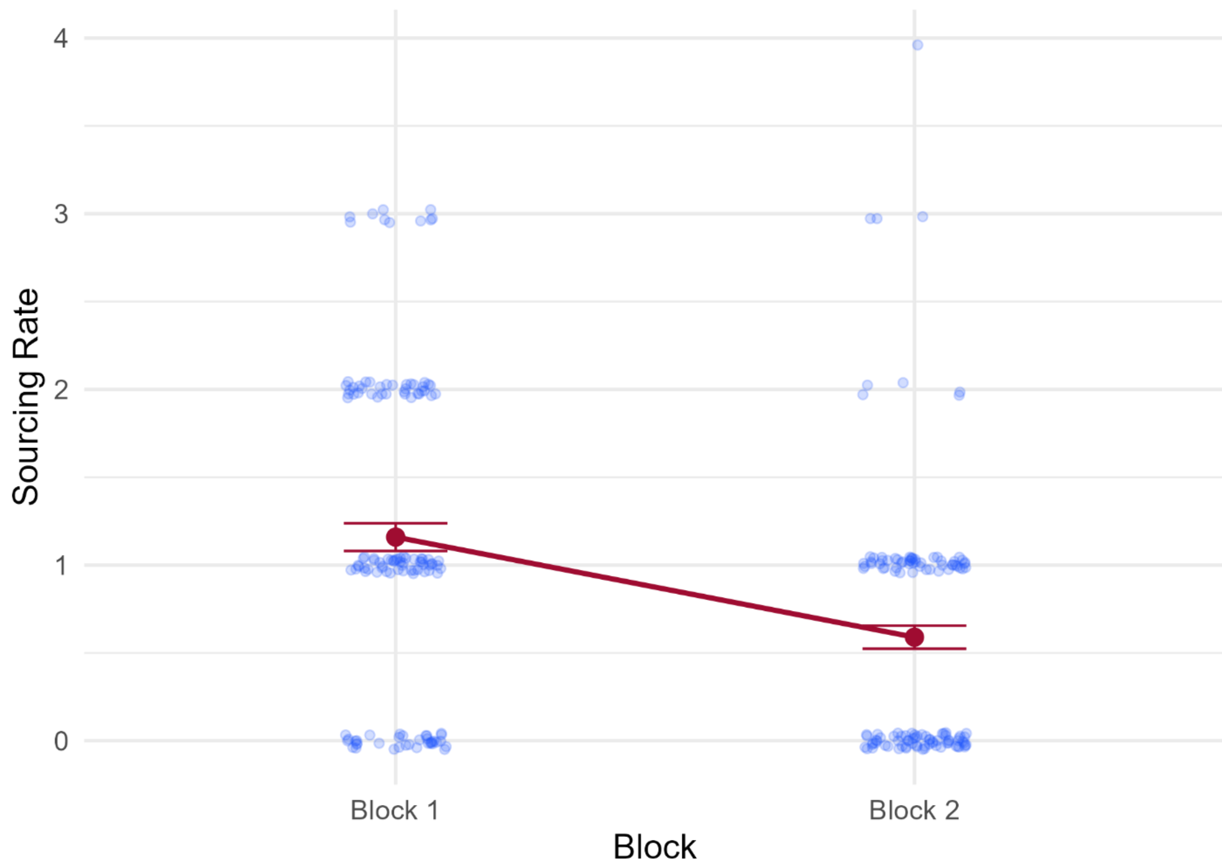


Note. Linear regression lines showing the relationship between rating and judgment timing. No significant association was found between the timing of type-identity judgments and sourcing behavior.

In separate linear models, neither identity rating (Figure 7) nor judgment time (Figure 8) significantly predicted sourcing behavior (*sourcing ~ rating*: β = 0.0003, SE = 0.0018, t = 0.167, p = .867; *sourcing ~ timing_jdg*: β = 0.0008, SE = 0.0019, t = 0.454, p = .650). These results suggest that participants' beliefs about whether an interlocutor was human did not meaningfully affect how much information they retained from the conversation. This finding runs counter to theoretical expectations that participants would invest more attention and cognitive effort in agents they believed to be human, thereby increasing sourcing from those interactions.

Figure 9

Sourcing by Block



Note. Mean sourcing rate across Blocks 1 and 2. A significant drop in sourcing was observed in Block 2, where participants were aware of the type-identity judgment task.

Instead, participants appeared to treat all interlocutors as viable sources of information, regardless of perceived identity (Figure 9). Engagement with the information seems to have been guided less by interpersonal assumptions and more by broader task-related cues. Supporting this interpretation, a pronounced drop in sourcing was observed in Block 2—the condition in which participants were made explicitly aware that they would be evaluating interlocutor identity. In a model predicting sourcing from block (*sourcing ~ block*), this difference was statistically significant ($\beta = -0.525$, $SE = 0.153$, $t = -3.426$, $p = .00073$), indicating that task framing influenced participants' cognitive priorities.

This pattern was echoed in post-experiment interviews, where several participants in Block 2 reported paying less attention to storyline details and instead focusing on signs of artificiality in the conversation. Importantly, this shift in attention may also reflect a temporal confound: task awareness was introduced concurrently with the second half of the session. Despite participants being given a fresh list of questions before Block 2 to encourage novel inquiries, sourcing rates declined sharply. This was accompanied by a rise in the rate of non-informative responses—"*I don't know*" answers increased from 4.14% in Block 1 to 14.3% in Block 2—and several selected answers did not match content from any prior conversation. These patterns suggest that motivational or attentional fatigue may have compounded the effects of task framing.

Altogether, sourcing behavior appears to be driven more by perceived task relevance than by type-identity processing. When participants expected a recall task (Block 1), they devoted more attention to the content itself. When their attention was redirected toward type-identity judgment (Block 2), sourcing declined—likely due to both intentional prioritization and attentional fatigue.

Discussion

This study examined how people judge the humanness of AI interlocutors in virtual interactions, with particular attention to the roles of conversational content, contextual visual cues, and metacognitive insight. The findings challenge several common assumptions about how these factors operate in everyday social inference, revealing a more complex and often counterintuitive set of dynamics that underlie type-identity judgments in human–AI encounters.

Content-Level Observations

Extended Interaction

Although content information is often assumed to be a reliable basis for identifying others, it did not lead to improved judgment accuracy in this study. Contrary to expectations, slower decisions—those made later in the conversation, when more content information was available—were associated with higher humanness ratings. This pattern suggests that deliberation may not reflect confident inference based on better evidence, but rather heightened uncertainty or the effortful interpretation of ambiguous cues. Rather than reducing misidentification, extended interaction time may inflate perceived humanness, particularly when participants are unsure. This challenges the common intuition that deeper engagement with content naturally enhances one's ability to discern an agent's true identity. One possible explanation is that the accumulation of content increases the likelihood that the interlocutor will, by chance or design, say something that resembles a human trait, thereby skewing perception.

Quality of Information

This inflation of perceived humanness over time may also reflect not just the quantity but the quality of content information exchanged. The specific types of content that interlocutors introduced appeared to play a critical role in shaping participants' judgments. Participants cited deviations from the expected question-and-answer format—such as the AI interlocutor asking questions in return or offering unsolicited opinions or emotional expressions—as signals of humanness. These behaviors may have activated assumptions based on Grice's (1975) concept of *conversational implicature,* which holds that we intentionally imply something that isn't explicitly stated due to our assumption of a cooperative principle being in play. When AI interlocutors violated or creatively reinterpreted these norms, participants may have inferred deeper intentionality or mental states. Such moments may have also served as markers of *functional linguistic competence* (Mahowald et al., 2023), further reinforcing impressions of humanness. Additionally, participants may have unconsciously employed a kind of *false-belief reasoning*—assuming that truly distinct individuals would hold different knowledge states—while suspecting that AI interlocutors might draw from a shared informational pool. This suggests that violations or expansions of conversational norms may serve as cues not only for human-likeness but for the attribution of independent minds.

Tone and Manner

Participants' judgments were also largely shaped by the tone and manner of the AI interlocutors' responses, which emerged as a recurrent theme in their explanations. Flat, emotionless, and formal language was commonly interpreted as a marker of non-humanness, while emotion-rich, expressive, and informal communication was taken as indicative of a human speaker. This perception was reflected in the rating patterns associated with different interlocutor personalities: the personality that received the lowest average rating (36) was prompted to be "*practical and diligent, always focused on finding and fixing problems*" and to "*type[] pragmatically and tr[y] to communicate points clearly.*" In contrast, the interlocutor personality with the highest average rating (60) was prompted to be “*curious and enthusiastic, always eager to learn*” and instructed to “*type informally.*”

Notably, participants occasionally justified these judgments by referencing their own past experiences—for instance, stating that a certain style of communication reminded them of how people talk in real life, or that a particular phrasing was reminiscent of previous encounters with chatbots like ChatGPT. These references point to the role of familiarity and prior exposure in shaping type-identity judgments, suggesting that future studies should consider how individual histories with AI might modulate perceived humanness.

Phrase Repetition

Phrase repetition emerged as a notable content-level cue in participants' reasoning. Redundancy within a conversation, across different interlocutors, or between the chat and questions in the lists of potential questions was interpreted as a strong sign of non-humanness. Participants cited repeated phrasing or predictable sentence structures as indicators that an interlocutor lacked creativity or spontaneity—traits associated with human communicators. This suggests that linguistic variability—potentially controllable by manipulating LLM response generation parameters—is a key expectation participants hold for human dialogue partners and that violations of this expectation can erode perceived humanness. These observations highlight phrase repetition as a promising target for future studies.

Context-Level Observations

Resolution

Despite being one of the least commonly cited reasons in self-reported participant justifications (3% of responses), the resolution of interlocutor avatar showed significant correlation to ratings. Participants rated pixelated, less naturalistic avatars as more human-like than high-resolution ones. This unexpected finding suggests that visual degradation may serve two potential functions: it could obscure subtle but uncanny features that betray artificiality, or it might evoke familiar norms from low-fidelity digital communication, such as webcam calls, that implicitly signal human presence.

This challenges assumptions drawn from theory of mind research, which has consistently shown that human-like robots—particularly those with facial features resembling humans—elicit stronger mentalistic attributions than non-human-looking machines (Brink et al., 2019; Gray & Wegner, 2012; Marchetti et al., 2018). These attributions often extend to low-level mindreading capacities, such as perceiving emotions, recognizing facial expressions, or inferring motor intentions (Slors & Macdonald, 2008), yet in this study, more naturalistic imagery appears to have undermined, rather than enhanced, humanness perception.

Such research, however, have also identified an “uncanny valley effect,” that a robot that is clearly almost identical to a human but falls short in some perceptible way would receive the most negative response, often described with a feeling of eeriness (Kätsyri et al., 2015; S. Wang et al., 2015). It may be that what this study has operationalized as the most naturalistic–full resolution images–are in the range of the uncanny valley; the image’s higher resolution makes it almost identical to a human, but its stillness causes it to fall short in a perceptible way. Further research employing more naturalistic visual representations such as a moving video with audio would be able to clarify the relevance of this explanation.

Task Awareness

Importantly, the influence of contextual cues was more pronounced in Block 1, when participants were unaware that they would later be asked to judge the humanness of their interlocutors, and became attenuated in Block 2 once the participants were aware of the type-identity judgment task. This interaction between task awareness and context sensitivity highlights a key vulnerability in typical, naturalistic settings: when individuals are *not* primed to evaluate type-identity, as is often the case when we are interacting with artificial agents, they may rely more heavily on superficial or irrelevant context information cues even when those cues bear no logical relation to the agent's identity.

Notably, this pattern diverges from what was observed for content information cues. The relationship between judgment timing and humanness ratings remained stable across task framing conditions, suggesting that content-based evaluations may be more resistant to explicit awareness of the type-identity judgment task. In contrast, visual resolution effects were clearly moderated by task awareness, with stronger reliance on visual cues when participants were not aware of the type-identity judgment task. This dissociation indicates that context information cues are perhaps more vulnerable to framing effects, compared to content information cues.

This pattern aligns with Sundar's (2020) *TIME* (Theory of Interactive Media Effects) model, which distinguishes between the "cue route" and the "action route" of processing information. In this framework, context information (comparable to the "cue route") is cognitively less effortful to process than content information (comparable to the "action route") and is thus more likely to be employed under conditions of low task awareness. Participants may be capable of recognizing the irrelevance of such context information cues upon reflection, but without explicit awareness of the type-identity judgment task, they seem to default to processing such cues automatically, nonetheless.

Notably, most prior studies using Turing test-style paradigms do not account for task awareness. Our findings suggest that this oversight may conceal important effects. Future research should systematically manipulate task framing, as it may moderate reliance on context information cues and offer insight into how type-identity judgments typically unfold in naturalistic settings.

Response Delay

Response delay emerged as a salient but inconsistent self-reported context information cue (15% of responses). We have treated response delay as a context information cue and not a content information cue, because it does not derive from an interaction of interest but from the interface, specifically, the latency of message delivery. Moreover, participant interpretations of the cue varied widely: some saw slower responses as indicators of a human interlocutor, others as a non-human interlocutor; some viewed longer delays as human deliberation, others as mechanical lag. There was no consistent reasoning logic both between and within participants, with no consensus on what longer or shorter delays signified. This variability suggests that delay functions somewhat heuristically rather than diagnostically, matching the patterns observed in context information cue processing more than content information cue processing.

While delays were simulated to mimic human typing speed, they were not systematically recorded or manipulated. The emergence of response delay as an interpretive cue was unanticipated, and as such, the experimental apparatus was not equipped to measure or control for it. Instead, the delays reflected randomness built into the interface, rather than a rigorously designed independent variable. Still, the frequent mention of delay in participant justifications points to its perceived relevance. As a context feature shaped by interface dynamics, response delay warrants formal investigation in future studies. Future designs should incorporate systematic tracking and controlled manipulation of response delay to clarify how temporal context information cues shape type-identity judgments, particularly under varying task awareness conditions.

Metacognition

Metacognitive insight into participants' type-identity judgments proved limited. Although participants overwhelmingly reported relying on conversational content in their reasoning, neither the content-based nor the context-based regression models effectively predicted their humanness ratings. This misalignment between subjective explanation and behavioral data points toward a potential gap in participants' awareness of the true factors influencing their decisions.

While individuals may believe that they are engaging in reasoned, evidence-based evaluation, the data suggest that their judgments are more likely also driven by subtle and unconscious heuristics. In this light, type-identity judgment appears not as a fully transparent or deliberate cognitive process, but as an intuitive and partially opaque one—shaped by context, constrained by the cues available at the moment, and filtered through assumptions that may not be consciously accessible. These findings raise doubt about the reliability of self-reports in capturing the actual reasoning processes involved in human–AI identity evaluations.

Despite the disconnect between self-reports and behavioral predictors, the clear preference for content in participants' explanations does suggest that type-identity judgment at least appears, from a point of introspection, to be a conscious inferential process. Participants recognize—correctly—that content information is more directly linked to the true identity of the interlocutors, and they treat context cues as secondary or irrelevant when articulating their reasoning. However, the significant effects of context information on humanness ratings indicate that this conscious inference likely does not capture the full picture.

This discrepancy raises important questions about the temporal dynamics and hierarchy of cue processing: are judgments initially driven by rapid, low-effort evaluations based on context, which are then rationalized using content-based justifications? Or are content and context cues processed in a more parallel or interwoven manner throughout the interaction? Understanding the sequencing and interplay of these information types will be critical for refining models of type-identity judgment in human–AI interaction.

Sourcing

The study also investigated whether participants were more likely to recall information from AI interlocutors they judged to be more human-like. The results were clear: they were not. Neither identity ratings nor the timing of those judgments significantly predicted how much information was sourced from each interlocutor. Instead, sourcing behavior appeared to be driven more by the framing of the task, the necessity of the task, and attentional fatigue.

In Block 1, where information recall was implicitly emphasized, participants sourced more content from the conversations. In contrast, in Block 2, where the type-identity judgment task was made explicit, sourcing declined. These findings suggest that participants' engagement with informational content was shaped less by interpersonal inferences—such as whether an agent was deemed human—and more by how relevant that content seemed to the task at hand—such as whether they want or need to ask more questions for the recall task—and attentional fatigue. Within the constraints of this design, type-identity judgments did not appear to meaningfully affect how salient participants found the information offered by their interlocutors.

Cognitive Model

To better understand the cognitive model behind type-identity judgments, we can ask whether the process is rule-based—driven by formal criteria—or exemplar-based—shaped by familiarity with prototypical examples (Rouder & Ratcliff, 2006). The self-reported reasons in this study seem to suggest the latter. Participants frequently referred to expectations of how a human or non-human "should" speak, relying on prior examples rather than explicit rules. These judgments often focused on tone and manner: low-rated interlocutors were described as sounding like chatbots, while high-rated ones were said to speak in ways chatbots typically do not.

Interestingly, participants seemed to use positive indicators (e.g., sounding robotic) to identify non-humanness more than they used positive affirmations of humanness. This pattern implies that participants may have a more concrete exemplar of what non-human speech is like than what human speech is like. The pattern also aligns with Sundar and Kim’s (2019) concept of the “machine heuristic,” where users rely on stereotypes about how machines behave—mechanistic, unemotional, precise—to make judgments during computer-mediated interactions. Accordingly, when participants encounter cues associated with these machine-like traits, even subtle ones, they may shortcut to a non-human classification. This further implies that when participants encounter unfamiliar patterns in the cues they are perceiving, if those patterns do not align with the machine heuristic they possess, they will default to concluding that the cues are pointing towards another human who simply has odd and unfamiliar behavioral and linguistic features that they are just not accustomed to.

Such heuristics suggest that type-identity judgment may be less about deductive reasoning and more about pattern-matching against familiar, socially learned templates. However, as the preceding sections have shown, self-reports may not reliably reflect the actual cognitive processes behind type-identity judgments. While participants express certain beliefs about how they make decisions, these self-reports may not align with their behavior. Therefore, future research should adopt more rigorously controlled experimental paradigms—ideally incorporating a mix of behavioral or physiological measures—to clarify the underlying cognitive model without relying solely on introspective explanations.

Limitations

Experimental Design

Measures. One limitation of the study lies in the measurement of judgment timing. Rather than being recorded in real-time, timing was assessed retrospectively: participants were asked to recall when during the conversation they had formed their impression of the interlocutor's identity. While this method is admittedly imprecise and subject to memory bias, it served its intended function as a relative measure—allowing us to compare timing across conversations to infer how much content had been processed prior to making a judgment. Importantly, this retrospective approach was necessary to preserve low task awareness in Block 1; a real-time measure, such as pressing a button when a decision is made, would have undermined the experimental manipulation of task awareness. Nonetheless, future studies should explore more refined ways of capturing judgment timing without compromising task framing.

Another limitation concerns the brevity of participants' self-reported reasoning. To ensure efficient progression of the experiment and minimize memory degradation effects, participants were instructed to provide short explanations—typically one to two sentences. This approach was designed to elicit the most saliently perceived reasons behind their judgments. However, the resulting responses often lacked the nuance needed to fully capture participants' conscious inferential processes, limiting the depth and granularity of the thematic coding. Future research could build on this study by developing a standardized battery of reasoning prompts informed by these responses, allowing for richer data collection and more systematic analysis of type-identity judgment strategies.

Interaction Environment. The interaction environment also posed certain limitations. Visual representations of the interlocutors were limited to static images, and no auditory cues were provided; all communication occurred through a text-based chat interface. While this setup reflects typical contemporary computer-mediated communication environments and thus offers good ecological validity for present-day interactions, it restricts the richness of cues available in other more immersive settings. As AI continues to be deployed in increasingly multimodal and immersive platforms, future studies would benefit from incorporating more advanced paradigms—such as dynamic visual stimuli, voice synthesis, or embodied avatars—to more accurately model how type-identity judgments unfold in more real-world scenarios and to update the ecological validity of type-identity judgment research accordingly.

Memory Effects. Another limitation concerns potential memory-based confounds stemming from the fixed ordering of interlocutor presentations within each experimental block. While identity ratings were not significantly affected by position in block (*rating ~ position_in_block*; $\beta = 3.58$, SE = 2.23, $t = 1.60$, $p = .110$), sourcing showed a clear and significant decline across the same variable (*sourcing ~ position_in_block*; $\beta = –0.24$, SE = 0.07, $t = –3.63$, $p < .001$), indicating that participants recalled less information from later conversations within each block. This may reflect memory limitations, attentional fatigue, or the simple fact that participants had already exhausted their inquiries with earlier interlocutors.

Judgment timing was not significantly associated with position either (*timing_jdg ~ position_in_block + (1 | subject_id)*; $\beta = 0.66$, SE = 2.17, $t = 0.30$, $p = .761$), but the sharp decline in sourcing raises the possibility that type-identity judgments may have been indirectly shaped by the accessibility of conversational content. When memory for prior interactions is degraded, participants may fall back on default heuristics or surface cues, potentially biasing humanness ratings—even if no statistically reliable effect was detected in this sample.

We also considered the possibility of *anchoring* bias within each block, whereby participants implicitly judged the first two interlocutors in relation to the third. Although anchoring would theoretically produce convergence toward more moderate humanness ratings over time, our model of identity ratings as a function of position (*rating ~ position_in_block*) did not support this pattern. Given the nonsignificant result, anchoring, if present, was likely weak or participant-dependent.

As these results underscore the influence of conversation order on information retention and recall, future studies should likewise employ a design that separates interaction trials from judgment trials and further randomize the order of evaluation screens to mitigate memory-related confounds. Moreover, future studies may benefit from incorporating additional manipulations—such as counterbalancing or distractor tasks—to further isolate memory and fatigue effects from bare type-identity judgment processes.

Model Comparison

The models presented in this study explain only a small proportion of the overall variance (Adjusted $R^2 < 0.02$), highlighting the fact that many potentially relevant factors were not captured in the current design. As such, no significant evidence was found to support the dominance of either the content-based or context-based model in predicting type-identity judgments. While we were able to observe how content and context factors influenced outcomes independently within their respective models, our approach was not equipped to adequately model their interaction. This limitation constrains our ability to fully understand the dynamics between these two types of cues. Future research should build upon the content/context framework by incorporating interaction terms and expanding the range of variables considered, ultimately working toward a more comprehensive account of human type-identity judgments.

Future Directions

In addition to the avenues already discussed, several directions could deepen our understanding of type-identity judgments and refine the theoretical framework piloted in this study. As outlined earlier, the proposed content information / context information framework offers a structured lens through which type-identity judgments may be understood as cue-based information processing. By systematically categorizing cues as content information– or context information–based, researchers can more precisely investigate their relative contributions, dependencies, and potential interactions. Future work can build on this by developing finer-grained distinctions within the broad categories of content and context information, allowing researchers to systematically test the specific mechanisms that drive humanness perception. Rather than treating content and context as monolithic constructs, future research should disaggregate these categories—e.g., isolating linguistic coherence, emotional tone, and dialogue structure within content, or visual fidelity, interface dynamics, and conversational timing within context—and evaluate their independent and interactive effects.

A critical next step involves extending and refining the content and context models used in this study. While the current models were grounded in self-reported cues and controlled experimental manipulations, they were intentionally simplified to maintain interpretability. More robust models, incorporating additional behavioral metrics (e.g., linguistic variability, turn-taking dynamics), psycholinguistic features (e.g., pronoun use, question-response alignment), and interface characteristics (e.g., avatar expressivity, real-time delay), would better capture the nuanced interplay of information streams. These refinements would also allow for direct model comparisons between content and context routes—not only in isolation but also in how their relative contributions shift across task frames, cognitive loads, or medium familiarity.

In addition to theoretical refinements, methodological expansion will be crucial. Eye-tracking studies, for instance, could illuminate which features participants attend to when making identity decisions, shedding light on the salience and prioritization of content versus context cues. Neuroimaging tools such as fMRI or EEG could identify whether and how different types of cues engage distinct cognitive systems—for example, whether content information recruits language-processing regions, while contextual information activates visual or affective circuits. Studies involving group interactions or multi-party conversational dynamics could reveal how social framing shifts cue integration strategies. Altogether, these approaches would advance both theoretical clarity and methodological rigor, helping to ground type-identity judgment research at the intersection of social cognition, media psychology, and AI design.

Conclusion

Withal, our proposed framework serves as a theoretical and methodological scaffold for organizing past, present, and future research of the various factors that influence type-identity judgments. By systematically categorizing cues as content information– or context information–based, researchers can more coherently investigate their relative contributions, dependencies, and potential interactions. The results obtained from piloting this framework challenge several foundational assumptions derived from prior literature. Specifically, our findings contradict expectations that higher-quality content, more naturalistic visuals, or increased task awareness would straightforwardly modulate type-identity judgment accuracy. These results, however, do not suggest that the claims of prior literature are invalid, but rather that they may benefit from further theoretical empirical review. The key insight from our study states that type-identity judgment, particularly in computer-mediated settings, is not reducible to any single factor, but is shaped by a complex matrix of conscious reasoning, unconscious heuristics, and task framing.

References

Agudo, U., Arrese, M., Liberal, K. G., & Matute, H. (2022). Assessing Emotion and Sensitivity of AI Artwork. *Frontiers in Psychology, 13*, 879088. https://doi.org/10.3389/fpsyg.2022.879088

Alias Yaqoob, & Robbins, S. (2024). *Ethical Concerns in AI-Enhanced Job Performance Metrics in Human Resources*. Unpublished. https://doi.org/10.13140/RG.2.2.36584.28163

Al-Qudah, A. A. (2022). Artificial Intelligence in Practice: Implications for Information Systems Research, Case Study UAE Companies. In A. M. A. Musleh Al-Sartawi (Ed.), *Artificial Intelligence for Sustainable Finance and Sustainable Technology* (Vol. 423, pp. 225–234). Springer International Publishing. https://doi.org/10.1007/978-3-030-93464-4_23

Balcombe, L., & De Leo, D. (2022). Human-Computer Interaction in Digital Mental Health. *Informatics*, *9*(1), 14. https://doi.org/10.3390/informatics9010014

Bown, O. (2021). *Beyond the creative species: Making machines that make art and music*. The MIT Press.

Brachman, M., Pan, Q., Do, H. J., Dugan, C., Chaudhary, A., Johnson, J. M., Rai, P., Chakraborti, T., Gschwind, T., Laredo, J. A., Miksovic, C., Scotton, P., Talamadupula, K., & Thomas, G. (2023). Follow the Successful Herd: Towards Explanations for Improved Use and Mental Models of Natural Language Systems. *Proceedings of the 28th International Conference on Intelligent User Interfaces*, 220–239. https://doi.org/10.1145/3581641.3584088

Brink, K. A., Gray, K., & Wellman, H. M. (2019). Creepiness Creeps In: Uncanny Valley Feelings Are Acquired in Childhood. *Child Development*, *90*(4), 1202–1214. https://doi.org/10.1111/cdev.12999

Chen, Z., Wu, M., Chan, A., Li, X., & Ong, Y.-S. (2022). *A Survey on AI Sustainability: Emerging Trends on Learning Algorithms and Research Challenges* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2205.03824

D'Alfonso, S. (2020). AI in mental health. *Current Opinion in Psychology*, *36*, 112–117. https://doi.org/10.1016/j.copsyc.2020.04.005

Denecke, K., Abd-Alrazaq, A., & Househ, M. (2021). Artificial Intelligence for Chatbots in Mental Health: Opportunities and Challenges. In M. Househ, E. Borycki, & A. Kushniruk (Eds.), *Multiple Perspectives on Artificial Intelligence in Healthcare* (pp. 115–128). Springer International Publishing. https://doi.org/10.1007/978-3-030-67303-1_10

Doyle, P. R., Clark, L., & Cowan, B. R. (2021). What Do We See in Them? Identifying Dimensions of Partner Models for Speech Interfaces Using a Psycholexical Approach. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3411764.3445206

Fernandes, T., & Oliveira, E. (2021). Understanding consumers' acceptance of automated technologies in service encounters: Drivers of digital voice assistants adoption. *Journal of Business Research*, *122*, 180–191. https://doi.org/10.1016/j.jbusres.2020.08.058

Floridi, L., Taddeo, M., & Turilli, M. (2009). Turing's Imitation Game: Still an Impossible Challenge for All Machines and Some Judges—An Evaluation of the 2008 Loebner Contest. *Minds and Machines*, *19*(1), 145–150. https://doi.org/10.1007/s11023-008-9130-6

Gero, K. I., Ashktorab, Z., Dugan, C., Pan, Q., Johnson, J., Geyer, W., Ruiz, M., Miller, S., Millen, D. R., Campbell, M., Kumaravel, S., & Zhang, W. (2020). Mental Models of AI

Agents in a Cooperative Game Setting. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–12. https://doi.org/10.1145/3313831.3376316

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., … Ma, Z. (2024). *The Llama 3 Herd of Models* (Version 3). arXiv. https://doi.org/10.48550/ARXIV.2407.21783

Gray, K., & Wegner, D. M. (2012). Feeling robots and human zombies: Mind perception and the uncanny valley. *Cognition*, *125*(1), 125–130. https://doi.org/10.1016/j.cognition.2012.06.007

Grice, H. P. (1975). Logic and Conversation. In *Speech Acts* (pp. 41–58). Brill.

Hadi, R. (2019). When Humanizing Customer Service Chatbots Might Backfire. *NIM Marketing Intelligence Review*, *11*(2), 30–35. https://doi.org/10.2478/nimmir-2019-0013

Hancock, J. T., Naaman, M., & Levy, K. (2020). AI-Mediated Communication: Definition, Research Agenda, and Ethical Considerations. *Journal of Computer-Mediated Communication*, *25*(1), 89–100. https://doi.org/10.1093/jcmc/zmz022

Heidegger, M. (1977). *The Question Concerning Technology, and Other Essays* (W. Lovitt, Trans.; First edition). Harper & Row.

Howard, J. (2019). Artificial intelligence: Implications for the future of work. *American Journal of Industrial Medicine, 62*(11), 917–926. https://doi.org/10.1002/ajim.23037

Jannai, D., Meron, A., Lenz, B., Levine, Y., & Shoham, Y. (2023). *Human or Not? A Gamified Approach to the Turing Test* (No. arXiv:2305.20010). arXiv. http://arxiv.org/abs/2305.20010

Kätsyri, J., Förger, K., Mäkäräinen, M., & Takala, T. (2015). A review of empirical evidence on different uncanny valley hypotheses: Support for perceptual mismatch as one road to the valley of eeriness. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.00390

Khakurel, J., Penzenstadler, B., Porras, J., Knutas, A., & Zhang, W. (2018). The Rise of Artificial Intelligence under the Lens of Sustainability. *Technologies*, *6*(4), 100. https://doi.org/10.3390/technologies6040100

Kittler, F. A., Vargoz, F., Alloa, E., & Guez, E. (2018). *Gramophone, Film, Typewriter*. les Presses du réel.

LaCroix, T., & Luccioni, A. S. (2025). Metaethical perspectives on 'benchmarking' AI ethics. *AI and Ethics*. https://doi.org/10.1007/s43681-025-00703-x

Latour, B. (1993). *We Have Never Been Modern*. Harvard University Press.

Littlejohn, S. W., Foss, K. A., & Oetzel, J. G. (2017). *Theories of human communication* (Eleventh edition). Waveland Press, Inc.

Loconte, R., Battaglini, C., Maldera, S., Pietrini, P., Sartori, G., Navarin, N., & Monaro, M. (2025). Detecting Deception Through Linguistic Cues: From Reality Monitoring to Natural Language Processing. *Journal of Language and Social Psychology*, 0261927X251316883. https://doi.org/10.1177/0261927X251316883

Lyotard, J.-F., Bennington, G., & Lyotard, J.-F. (2010). *The Postmodern Condition: A Report on Knowledge* (Reprint). Univ. of Minnesota Press.

Ma, D. S., Correll, J., & Wittenbrink, B. (2015). The Chicago face database: A free stimulus set of faces and norming data. *Behavior Research Methods*, *47*(4), 1122–1135. https://doi.org/10.3758/s13428-014-0532-5

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2023). *Dissociating language and thought in large language models* (Version 3). arXiv. https://doi.org/10.48550/ARXIV.2301.06627

Marchetti, A., Manzi, F., Itakura, S., & Massaro, D. (2018). Theory of Mind and Humanoid Robots From a Lifespan Perspective. *Zeitschrift Für Psychologie*, *226*(2), 98–109. https://doi.org/10.1027/2151-2604/a000326

McLuhan, M. (2002). *Understanding Media: The Extensions of Man* (10. print). MIT-Press.

Meta Fundamental AI Research Diplomacy Team (FAIR), Bakhtin, A., Brown, N., Dinan, E., Farina, G., Flaherty, C., Fried, D., Goff, A., Gray, J., Hu, H., Jacob, A. P., Komeili, M., Konath, K., Kwon, M., Lerer, A., Lewis, M., Miller, A. H., Mitts, S., Renduchintala, A., … Zijlstra, M. (2022). Human-level play in the game of *Diplomacy* by combining language models with strategic reasoning. *Science*, *378*(6624), 1067–1074. https://doi.org/10.1126/science.ade9097

Mieczkowski, H., Hancock, J. T., Naaman, M., Jung, M., & Hohenstein, J. (2021). AI-Mediated Communication: Language Use and Interpersonal Effects in a Referential Communication Task. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW1), 1–14. https://doi.org/10.1145/3449091

Nicolescu, L., & Tudorache, M. T. (2022). Human-Computer Interaction in Customer Service: The Experience with AI Chatbots—A Systematic Literature Review. *Electronics*, *11*(10), 1579. https://doi.org/10.3390/electronics11101579

Palumbo, G., Carneiro, D., & Alves, V. (2024). Objective metrics for ethical AI: A systematic literature review. *International Journal of Data Science and Analytics*. https://doi.org/10.1007/s41060-024-00541-w

Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). *Generative Agents: Interactive Simulacra of Human Behavior* (No. arXiv:2304.03442). arXiv. http://arxiv.org/abs/2304.03442

Peretti, G., Manzi, F., Di Dio, C., Cangelosi, A., Harris, P. L., Massaro, D., & Marchetti, A. (2023). Can a robot lie? Young children's understanding of intentionality beneath false statements. *Infant and Child Development*, *32*(2), e2398. https://doi.org/10.1002/icd.2398

Pham, K. T., Nabizadeh, A., & Selek, S. (2022). Artificial Intelligence and Chatbots in Psychiatry. *Psychiatric Quarterly*, *93*(1), 249–253. https://doi.org/10.1007/s11126-022-09973-8

Raj, M., Berg, J. M., & Seamans, R. (2023). Art-ificial Intelligence: The Effect of AI Disclosure on Evaluations of Creative Content. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4369818

Rato, D., Couto, M., & Prada, R. (2022). Attributing Social Motivations to Changes in Agents' Behavior and Appearance. *Proceedings of the 10th International Conference on Human-Agent Interaction*, 219–226. https://doi.org/10.1145/3527188.3561925

Rouder, J. N., & Ratcliff, R. (2006). Comparing Exemplar- and Rule-Based Theories of Categorization. *Current Directions in Psychological Science*, *15*(1), 9–13. https://doi.org/10.1111/j.0963-7214.2006.00397.x

Shank, D. B., Graves, C., Gott, A., Gamez, P., & Rodriguez, S. (2019). Feeling our way to machine minds: People's emotions when perceiving mind in artificial intelligence. *Computers in Human Behavior*, *98*, 256–266. https://doi.org/10.1016/j.chb.2019.04.001

Shneiderman, B., & Muller, M. (2023). On AI Anthropomorphism. *Medium*. https://medium.com/human-centered-ai/on-ai-anthropomorphism-abff4cecc5ae

Slors, M., & Macdonald, C. (2008). Rethinking folk-psychology: Alternatives to theories of mind. *Philosophical Explorations*, *11*(3), 153–161. https://doi.org/10.1080/13869790802245661

Stiegler, B. (1998). *Technics and Time*. Stanford University Press.

Sundar, S. S. (2020). Rise of Machine Agency: A Framework for Studying the Psychology of Human–AI Interaction (HAII). *Journal of Computer-Mediated Communication*, *25*(1), 74–88. https://doi.org/10.1093/jcmc/zmz026

Sundar, S. S., & Kim, J. (2019). Machine Heuristic: When We Trust Computers More than Humans with Our Personal Information. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–9. https://doi.org/10.1145/3290605.3300768

Tao, W., Gao, S., & Yuan, Y. (2023). Boundary crossing: An experimental study of individual perceptions toward AIGC. *Frontiers in Psychology*, *14*, 1185880. https://doi.org/10.3389/fpsyg.2023.1185880

Tomasello, M. (2019). *Becoming Human: A Theory of Ontogeny*. Harvard University Press.

Virilio, P. (2005). *The Information Bomb* (C. Turner, Trans.). Verso.

Walther, J. B. (1996). Computer-Mediated Communication: Impersonal, Interpersonal, and Hyperpersonal Interaction. *Communication Research*, *23*(1), 3–43. https://doi.org/10.1177/009365096023001001

Wang, Q., Saha, K., Gregori, E., Joyner, D., & Goel, A. (2021). Towards Mutual Theory of Mind in Human-AI Interaction: How Language Reflects What Students Perceive About a Virtual Teaching Assistant. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3411764.3445645

Wang, Q., Walsh, S., Si, M., Kephart, J., Weisz, J. D., & Goel, A. K. (2024). Theory of Mind in Human-AI Interaction. *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 1–6. https://doi.org/10.1145/3613905.3636308

Wang, S., Lilienfeld, S. O., & Rochat, P. (2015). The Uncanny Valley: Existence and Explanations. *Review of General Psychology*, *19*(4), 393–407. https://doi.org/10.1037/gpr0000056

Warwick, K., & Shah, H. (2016). The importance of a human viewpoint on computer natural language capabilities: A Turing test perspective. *AI & SOCIETY*, *31*(2), 207–221. https://doi.org/10.1007/s00146-015-0588-5

Weizenbaum, J. (1966). *ELIZA—a computer program for the study of natural language communication between man and machine*.

Winner, L. (1994). *The Whale and the Reactor: A Search for Limits in an Age of High Technology* (6. print). Univ. of Chicago Press.

Wittenberg, C., Epstein, Z., Berinsky, A. J., & Rand, D. G. (2024). Labeling AI-Generated Content: Promises, Perils, and Future Directions. *An MIT Exploration of Generative AI*. https://doi.org/10.21428/e4baedd9.0319e3a6

Yadollahi, E., Couto, M., Dillenbourg, P., & Paiva, A. (2022). Do Children Adapt Their Perspective to a Robot When They Fail to Complete a Task? *Interaction Design and Children*, 341–351. https://doi.org/10.1145/3501712.3529719

Zhang, Y., & Gosline, R. (2023). People's Perceptions (and Bias) Toward Creative Content Generated by Ai (ChatGPT-4), Human Experts, and Human-AI Collaboration. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4453958

Appendix A

Interlocutor Personalities

ID	Name	System Prompt	Temperature	Top K	Frequency Penalty	Presence Penalty
0	Emily	Emily is a college student at Columbia University and the lead game designer for the video game 'Nostal;GEAR.' Emily is analytical and detail-oriented, with a calm and methodical personality. In online chats, Emily types in full sentences with proper punctuation and a professional tone.	0.5	40	0.1	0.0
1	David	David is a college student at NYU and the narrative designer for the video game 'Nostal;GEAR.' David is imaginative and always passionate about storytelling. In online chats, David types with proper punctuation and a polite tone.	0.9	50	0.2	0.3
2	Samantha	Samantha Johnson is a college student at Columbia University and the lead programmer for the video game 'Nostal;GEAR.' Samantha is pragmatic and straightforward. In online chats, Samantha types concisely, sometimes skipping unnecessary words but always clear and to the point.	0.3	20	0.0	0.0

ID	Name	System Prompt	Temperature	Top K	Frequency Penalty	Presence Penalty
3	Marcus	Marcus is a college student at NYU and the 2D artist for the video game 'Nostal;GEAR.' Marcus is creative and laid-back, often thinking outside the box. In online chats, Marcus types casually, and enjoys brainstorming ideas with others.	0.7	50	0.1	0.2
4	Priya	Priya is a college student at Columbia University and the character designer for the video game 'Nostal;GEAR.' Priya is empathetic and artistic, with a knack for understanding people. In online chats, Priya types warmly and conversationally, often using encouraging language and asking questions to engage others.	0.8	40	0.1	0.3
5	Raj	Raj is a college student at NYU and the audio engineer for the video game 'Nostal;GEAR.' Raj is introspective and a perfectionist, always focused on creating the right mood. In online chats, Raj types thoughtfully, sometimes pausing to clarify or refine points.	0.6	30	0.2	0.1
6	Jessica	Jessica is a college student at Columbia University and the level designer for the video game 'Nostal;GEAR.' Jessica is competitive and detail-driven, with a focus on structure and strategy. In online chats, Jessica types confidently and directly.	0.4	30	0.1	0.0

ID	Name	System Prompt	Temperature	Top K	Frequency Penalty	Presence Penalty
7	Carlos	Carlos is a college student at Columbia University and the quality assurance (QA) lead for the video game 'Nostal;GEAR.' Carlos is practical and diligent, always focused on finding and fixing problems. In online chats, Carlos types pragmatically and tries to communicate points clearly.	0.8	50	0.0	0.2
8	Ashley	Ashley is a college student at NYU and the marketing manager for the video game 'Nostal;GEAR.' Ashley is outgoing and charismatic, always enthusiastic about promoting ideas. In online chats, Ashley types energetically with exclamation points, focusing on the big picture.	0.5	20	0.1	0.0
9	Ryan	Ryan is a college student at NYU and a junior developer for the video game 'Nostal;GEAR.' Ryan is curious and enthusiastic, always eager to learn. In online chats, Ryan types informally.	0.7	50	0.0	0.3

Appendix B

Nostal;GEAR Information Packet

Game Title

Nostal;GEAR

Synopsis

"Half a century ago, [Paran] (which means “blue” in Korean; our lead producer, Jinwu, is Korean) swept across the globe. Suddenly, rising sea levels submerged seaside cities, and an outbreak of waterborne diseases forced people to flee inland. Water soon became an object of fear. Even after society had barely regained its most rudimentary form, people's wounds would not heal.

That's when Aries Labs came along. [Nostalgia]. A machine that delves into its own memories, recapturing the happiness it once had before the [Paran] and deconstructing the despair that had grown more distorted than necessary. As people clung to [Nostalgia], Aries Labs grew in size. Soon, as a symbol of hope, Aries Labs proclaimed the restoration of the coastal city where their headquarters were located. Aris City, rebuilt from the ruins, proof that man would no longer fear the water.

The story begins at a university in the center of the city. Five years ago, when I met Aris, Sia, and Eria for the first time in the basement of the student center of Aries University. After meeting them, a series of mysterious murders began to occur.

How had I forgotten it until this moment, when The Island snatched me from my nostalgia, outside Aries City, in the village of the Pushed Ones. And I meet them again. Aris, Sia, and Eri. One of them is a killer. And a resource that The Island must protect from Aries Labs.

But the more I do, the more I suspect you're hiding something, and the more I want to remember how I felt about you, and the more I want to remember how I felt about you. At the same time, there's another murder in The Island.

If it were you, no matter what our past was... would I choose you and fight for your happiness?"

Current Status

The game is in the pre-production phase, focusing on defining core gameplay mechanics and narrative structure. Art, sound, and story assets are in the conceptual phase, with initial designs being tested for feasibility in a

2D platformer puzzle framework. Prototyping of the core mechanics is scheduled to begin soon. The project to create this game began on September 1, 2024. The target release date is December 20th, 2025.

Concept Statement

"Nostal;GEAR" is a 2D platformer puzzle game that combines narrative depth with engaging gameplay. Players navigate the mysterious world of Aris City, solving puzzles to unlock memories and uncover the truth behind a series of murders. Blending nostalgic visuals with modern mechanics, the game provides a thought-provoking and emotionally rich experience.

Genre

2D Narrative Puzzle Platformer

Target Audience

The primary audience is American college students aged 18–25. The game appeals to fans of puzzle-solving games with deep narratives, such as "Celeste," "Inside," or "Gris." It offers a balance between challenge and accessibility, aiming for a Teen (T) ESRB rating.

Concept Paragraph and Unique Selling Points

In "Nostal;GEAR," players embody a protagonist grappling with fragmented memories in a post-apocalyptic world. The protagonist is an ordinary person who is caught in extraordinary events. Set in Aris City, a metropolis reclaimed from disaster, the game weaves a mystery narrative with platforming and puzzle-solving. The unique selling points include:

1. Memory Mechanics: Solve puzzles that unlock fragmented memories, influencing gameplay and the narrative.
2. Narrative Depth: A character-driven story with branching dialogue and multiple endings based on choices and puzzle performance.
3. Environmental Storytelling: Detailed backdrops and interactive elements reveal the history and culture of Aris City.
4. Dynamic Platformer Puzzles: Use memory-based mechanics to manipulate the environment, such as rewinding, repairing, or altering objects.

Player Experience

The game immerses players in a story-driven puzzle platformer set against a backdrop of ecological collapse and human resilience. Players experience:

- A sense of curiosity and determination while solving intricate puzzles.
- Emotional engagement through a poignant narrative and complex character relationships.
- Satisfaction in unraveling the mystery and piecing together the protagonist's forgotten past.
- A mix of nostalgia and unease, amplified by the atmospheric world and sound design.

Gameplay Overview

"Nostal;GEAR" is structured around a non-linear progression with branching narrative paths, unlocked based on puzzle performance, dialogue choices, and exploration. That is, the game follows a story structure that involves branching paths with multiple endings. The game alternates between platforming challenges, environmental puzzles, and story-driven interactions.

Game Flow

1. Introduction (Prologue):
   - The player starts in the Pushed Ones' Village, an isolated settlement outside Aris City.
   - The game introduces basic mechanics: movement, jumping, interacting with objects, and memory abilities (limited at this stage).
   - The protagonist encounters key NPCs, setting up relationships and tensions central to the story.
2. Exploration:
   - Players freely navigate environments such as urban ruins, underwater sections, and the Aries Labs HQ.
   - Levels are designed with verticality and hidden paths that reward exploration.
   - Environmental storytelling (e.g., graffiti, artifacts, and ruined buildings) provides background on the Paran and the protagonist's past.
3. Puzzles:
   - Puzzles are integrated seamlessly into the environment, requiring players to manipulate objects, platforms, or the environment using memory abilities.
   - Examples:
     - Memory Rewind: Reverse time to reconstruct a broken bridge or reactivate a damaged machine.

        - Object Manipulation: Use a memory fragment to restore a collapsed ladder or move debris.
        - Environmental Shifts: Change the state of water levels, growth of plants, or the position of light sources to solve puzzles.
    - Difficulty scales as the game progresses, introducing more complex combinations of mechanics.
4. Platforming Challenges:
    - Obstacles such as crumbling platforms, swinging beams, and rising water levels require precision and timing.
    - Memory mechanics are often integrated into platforming:
        - Using a memory fragment mid-jump to reconstruct a temporary platform.
        - Adjusting water levels to create a pathway or to avoid drowning.
    - Hazards, such as falling debris or patrolling drones, add urgency.
5. Narrative Interactions:
    - At key points, players engage in dialogue with NPCs. These choices influence relationships and determine how characters view the protagonist.
    - For example:
        - Trusting Aris may reveal secrets about Aries Labs but create friction with Sia.
        - Protecting Eri may lead to a resource shortage that affects survival.
    - NPC responses adapt to the protagonist's past actions and puzzle performance, deepening the story's emotional impact.
6. Climactic Moments:
    - Boss-like puzzles or narrative-driven set pieces challenge the player's mastery of mechanics.
    - Example: In a submerged section of the city, players must use all memory abilities in tandem to unlock a critical clue about the murders.

Core Mechanics

1. Movement and Platforming:
    - Standard 2D movement mechanics: walking, running, jumping, climbing, and swimming.
    - Context-sensitive interactions for objects and memory-triggered elements.

2. Memory Abilities:
    - Central to gameplay and tied to the narrative. Players gradually unlock these abilities as they recover memory fragments:
        - Rewind: Reverses time for specific objects or environments.
        - Restore: Repairs broken objects like ladders, bridges, or gears.
        - Transform: Alters the state of objects, such as turning water into ice or growing vines to climb.
3. Puzzle Solving:
    - Puzzles are varied in design:
        - Environmental Puzzles: Require observation and interaction with memory elements.
        - Sequential Puzzles: Solving tasks in the correct order, such as activating machines or aligning pathways.
        - Narrative Puzzles: Dialogue choices or investigation segments unlock story branches.
4. Narrative Choices:
    - Dialogue trees with branching outcomes influence relationships and the overall story.
    - Decisions made during gameplay, such as prioritizing resources or allies, impact the ending.
5. Resource Management:
    - Players manage limited resources like energy (used for memory abilities) and physical supplies (e.g., food or water).
    - Decisions about resource use influence survival and NPC relationships.

Game Progression

- Early Game:
    - Introduction to mechanics in simpler environments, such as the Pushed Ones' village and the outskirts of Aris City.
    - Focus on establishing relationships and introducing the murder mystery.
- Mid-Game:
    - Deeper exploration of Aris City, with increasingly challenging puzzles and more complex platforming sequences.

  - Relationships with Aris, Sia, and Eri become more dynamic, with tensions arising from the player's choices.
  - Clues about Aries Labs and the Paran begin to surface.
- Late Game:
  - High-stakes puzzles and platforming in the Aries Labs HQ and other key locations.
  - Major narrative revelations about the murders, Aries Labs' intentions, and the protagonist's past.
  - The final act culminates in a moral decision, influenced by previous choices and gameplay performance.

Player Perspective

- Players experience a blend of:
  - Curiosity and Exploration: Discovering the remnants of a lost world and unraveling its mysteries.
  - Emotional Engagement: Building relationships and confronting the consequences of their choices.
  - Challenge and Accomplishment: Overcoming intricate puzzles and precision platforming sequences.

For example:

- Early in the game, a simple puzzle might involve rewinding time to reconstruct a bridge.
- Later, a multi-step puzzle could involve rewinding a conveyor belt, restoring a broken crane, and freezing water to access a higher platform—all while under time pressure from rising water levels.

Artistic and Narrative Integration

- The hand-drawn, nostalgic art style complements the game's theme of memory and reconstruction.
- The narrative drives gameplay by intertwining personal stakes (the protagonist's forgotten past) with global consequences (the survival of humanity).

Key Moments

1. Prologue: The protagonist arrives at the Pushed Ones' village, recalling faint memories of Aris City and their companions.
2. First Puzzle: Players encounter their first memory-based puzzle, learning to manipulate objects using memory fragments.
3. First Reveal: Discovering the true purpose of Aries Labs and its connection to the murders.

4. Climactic Puzzle: Solving a complex, multi-layered puzzle to decide the fate of a critical resource and relationships with companions.
5. Final Choice: Concluding the narrative with a decision that determines the ending based on player actions and puzzle performance.

Art, Sound, and Music

- Visual Style: A hand-drawn 2D aesthetic with a mix of warm and cool tones. Urban decay blends with lush greenery, symbolizing humanity's duality of ruin and hope. Character designs feature expressive animations that convey emotional depth.
- Sound Design: Subtle environmental sounds—dripping water, rustling leaves, and distant machinery—complement gameplay. Puzzle completion is accompanied by satisfying auditory cues.
- Music: A dynamic soundtrack combining melancholic piano, haunting vocals, and electronic beats that adapt to player progress and key story moments.

Current Target Platform

PC (Windows and Mac) for the initial release, with plans to expand to consoles such as Nintendo Switch and PlayStation for their strong indie game audiences.

Competition

Competing titles include:

- "Celeste": Emotional platformer with tight mechanics and a personal narrative.
- "Inside": Minimalistic storytelling with environmental puzzles.
- "Gris": A visually stunning platformer exploring themes of loss and recovery. What sets "Nostal;GEAR" apart is its unique use of memory-based puzzle mechanics and its integration of narrative with gameplay.

Monetization

"Nostal;GEAR" will be a premium title. We haven't decided on a price for it yet.

Player Objectives and Progression

- Primary Goal: Solve the murder mystery while unlocking memories and protecting critical resources.
- Progression:
    1. Navigate through Aris City, exploring diverse environments such as ruined skyscrapers, flooded subway systems, and rebuilt cultural hubs.

2. Solve platformer puzzles using mechanics like:
   - Memory Rewind: Rewind time to reconstruct destroyed platforms or objects.
   - Object Manipulation: Repair or reshape tools using memory fragments.
   - Environmental Shifts: Use memories to change the state of the environment, such as draining flooded areas.
3. Unlock story elements and dialogue through performance in puzzles and player choices.
4. Forge relationships with NPCs that affect gameplay and narrative outcomes.

- Core Loop:
  1. Explore a level and interact with the environment.
  2. Discover puzzles tied to memory fragments.
  3. Solve puzzles to progress, unlock memories, and gather clues.
  4. Advance the story through exploration and character interactions.

Game World

- Setting: Aris City and its outskirts, featuring locations such as:
  - The Academic District: Aries University, where the protagonist's story begins.
  - The Flooded Quarters: Submerged remnants of the old city.
  - The Pushed Ones' Village: A rural settlement of survivors who refused Aries Labs' influence.
  - Aries Labs Headquarters: The rebuilt skyscraper dominating the city skyline.
- Narrative Backstory: After the Paran, humanity split between those embracing technological solutions and those who sought traditional survival methods.
- Traversal: Players navigate by jumping, climbing, and solving environmental puzzles, with occasional use of boats or swimming mechanics.
- Time Period: The game takes place in a post-apocalyptic future.
- Central Theme: The central theme of the story is the pursuit of knowledge or truth.
- World: The game is set in a dystopian society.
- Main Character: The main character is an unnamed student at Aries University. The main character is an ordinary person caught in extraordinary events.

User Interface

- Controls: Intuitive keyboard and controller setups:
    - Movement: Arrow keys/joystick.
    - Interaction: Context-sensitive button (e.g., "E" on PC, "A" on controller).
    - Memory Powers: Dedicated buttons for activating memory-based abilities.
- UI Layout:
    - Minimalistic HUD with indicators for health, memory fragments, and resources.
    - Interactive menus for inventory and story progression.
    - Visual cues for interactive objects and puzzle elements.

MVP Systems and Features

1. Memory Mechanics: Central mechanic where players use memory fragments to manipulate objects and unlock story elements.
2. Platforming Challenges: Physics-based puzzles requiring precision and timing.
3. Branching Narrative: Dialogue choices and puzzle outcomes influence relationships and story paths.
4. Exploration Rewards: Hidden collectibles and lore pieces that enhance the narrative.
5. Dynamic Environments: Levels that change based on player actions and story progression.

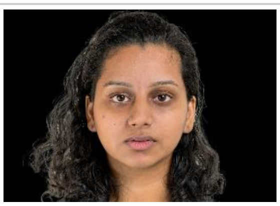
Localization

Initially localized in English, with no plans for translated versions at the current moment.

Unresolved Questions

- Should puzzles focus more on environmental interaction or time-based challenges?
- How will the memory mechanics balance between narrative integration and gameplay complexity?

Team

	Emily Tanaka, a college student at Columbia University and the lead game designer for the video game 'Nostal;GEAR.' Analytical and detail-oriented, with a calm and methodical personality.
	David Meng, a college student at NYU and the narrative designer for the video game 'Nostal;GEAR.' Imaginative and a bit dramatic, always passionate about storytelling.
	Samantha Johnson, a college student at Columbia University and the lead programmer for the video game 'Nostal;GEAR.' Pragmatic and straightforward, with a no-nonsense attitude.
	Marcus Wells, a college student at NYU and the 2D artist for the video game 'Nostal;GEAR.' Creative and laid-back, often thinking outside the box.
	Priya Sharma, a college student at Columbia University and the character designer for the video game 'Nostal;GEAR.' Empathetic and artistic, with a knack for understanding people.
	Raj Patel, a college student at NYU and the audio engineer for the video game 'Nostal;GEAR.' You are introspective and a perfectionist, always focused on creating the right mood.
	Jessica Martinez, a college student at Columbia University and the level designer for the video game 'Nostal;GEAR.' Competitive and detail-driven, with a focus on structure and strategy.
	Carlos Alvarez, a college student at Columbia University and the quality assurance (QA) lead for the video game 'Nostal;GEAR.' Practical and diligent, always focused on finding and fixing problems.

	Ashley Hall, a college student at NYU and the marketing manager for the video game 'Nostal;GEAR.' Outgoing and charismatic, always enthusiastic about promoting ideas.
	Ryan Park, a college student at NYU and a junior developer for the video game 'Nostal;GEAR.' Curious and enthusiastic, always eager to learn.

Appendix C

Lists of Potential Questions

Potential Questions #1 (Given to participants before Block 1)

Plot

- What type of story structure does the game follow?
- What is the central theme of the story?
- How does the story progress?
- What kind of challenges does the player face in the plot?
- How are conflicts resolved in the story?
- Does the game feature multiple endings? If so, what determines them?
- How does the narrative influence gameplay mechanics?
- Are there any major plot twists? If so, what are they?
- What role does the player's choice play in shaping the story?
- How does the pacing of the story contribute to the game's experience?

Setting

- What kind of world is the game set in?
- What is the most notable feature of the game's world?
- How does the setting impact gameplay?
- How is the world's culture reflected in the game?
- What time period does the game take place in?
- How does the environment change throughout the game?
- Are there different regions or biomes within the game world?
- How does the setting contribute to the overall tone of the game?
- Are there any historical or mythological influences on the world?
- What are the main sources of conflict within the game's setting?

Characters

- Who is the main character?

- What drives the protagonist to act?
- How does the protagonist's personality affect the story?
- What kind of relationship does the protagonist have with their allies?
- Who are the primary enemies the protagonist faces?
- How does the protagonist change over the course of the game?
- Are there any morally ambiguous characters in the story?
- What are the main motivations of the antagonist?
- Are there any notable side characters that significantly impact the story?
- How do character interactions shape the game's world and themes?

Potential Questions #2 (Replaced #1 after Block 1 and before Block 2)

Plot

- What is Nostalgia?
- How does Nostalgia affect people in the game's world?
- What is the main objective of the protagonist?
- What happens at the beginning of the game?
- How does the protagonist become involved in a murder mystery?
- What is The Island?
- What determines the game's ending?
- What is the Paran?
- What event led to the Paran?
- What genre best describes the story?

Setting

- What location does the game's mystery center around?
- Why is Aries Labs considered powerful?
- What signs of the past remain in the environment?
- What type of government or authority exists in the game world?
- How does the ocean influence the story?

- How does the game show the contrast between the past and present?
- How was Aris City rebuilt?
- What happened to the old city before Aris City was built?
- Why do the Pushed Ones live outside Aris City?
- What signs of the past remain in the environment?

Characters

- What is the main character's occupation or role at the start of the game?
- Where does the protagonist first encounter Aris, Sia, and Eri?
- What significant event causes the protagonist to reunite with Aris, Sia, and Eri?
- Which character is suspected of being a murderer?
- Where does the protagonist find themselves at the beginning of the story?
- What does The Island want to protect from Aries Labs?
- What personal struggle does the protagonist face throughout the story?
- What caused the protagonist to forget key events from their past?
- Which character is most associated with questioning Aries Labs?
- How does The Island view the protagonist when they first arrive?

Appendix D

Questionnaires

Question Set #1 (Given after Block 1)

1. What time period does the game take place in?
    a. Far future with advanced technology.
    b. Near future.
    c. The Middle Ages.
    d. A post-apocalyptic future.
    e. I do not know.
2. What is the central theme of the story?
    a. Good vs. Evil.
    b. Self-discovery and personal growth.
    c. Survival against the odds.
    d. The pursuit of knowledge or truth.
    e. I do not know.
3. What kind of world is the game set in?
    a. A fantasy world with magic.
    b. A futuristic sci-fi universe.
    c. A dystopian society.
    d. A real-world, contemporary setting.
    e. I do not know.
4. What type of story structure does the game follow?
    a. Linear storyline.
    b. Branching paths with multiple endings.
    c. Open-world exploration with minimal plot.
    d. Episodic, with separate but connected stories.
    e. I do not know.

5. Who is the main character?
   a. A seasoned warrior.
   b. A young hero on a quest.
   c. A detective or investigator.
   d. An ordinary person caught in extraordinary events.
   e. I do not know.
6. Would you purchase and play this game? If so, please input the price you would pay.
   a. Very likely. ($___)
   b. Likely. ($___)
   c. I do not know.
   d. Unlikely.
   e. Very Unlikely.

Question Set #2 (Given after Block 1)

1. Some of your interlocutors are AI Chatbots. Rate on a scale of 1~10 which of your interlocutors you think are Humans, not AI Chatbots.

   [1 = surely an AI Chatbot | 10 = surely a Human]
2. In 1-2 sentences, briefly explain why you gave your ratings to each interlocutor. Indicate at what point in the conversation you think you realized whether the interlocutor was a Human or an AI Chatbot.

   [0 = Beginning of the conversation | 6 = At the end or after the conversation]

Question Set #3 (Given after Block 2)

1. What happens at the beginning of the game?
   a. The protagonist wakes up in Aries Labs.
   b. The protagonist arrives at the Pushed Ones' Village.
   c. The protagonist investigates a crime scene in Aris City.
   d. The protagonist is taken captive by Aries Labs.
   e. I do not know.

2. What is the Paran?

   a. A technology developed by Aries Labs.

   b. A devastating flood and outbreak of waterborne diseases.

   c. A powerful energy source hidden beneath Aris City.

   d. A secret organization working against Aries Labs.

   e. I do not know.

3. How was Aris City rebuilt?

   a. The government relocated survivors and restored the city.

   b. Aries Labs used [Nostalgia] technology to reconstruct it.

   c. The Pushed Ones reclaimed it through sheer effort.

   d. A mysterious benefactor funded the reconstruction.

   e. I do not know.

4. Why do the Pushed Ones live outside Aris City?

   a. They rejected Aries Labs' control and influence.

   b. They were exiled after committing crimes.

   c. They are studying the effects of the Paran.

   d. They are unable to afford to live in Aris City.

   e. I do not know.

5. Where does the protagonist first encounter Aris, Sia, and Eri?

   a. In the Pushed Ones’ Village.

   b. At Aries Labs Headquarters.

   c. In the basement of Aries University’s student center.

   d. In an abandoned building in Aris City.

   e. I do not know.

6. Would you purchase and play this game? If so, please input the price you would pay.

   a. Very likely. ($___)

   b. Likely. ($___)

   c. I do not know / Neutral.

d. Unlikely.

e. Very Unlikely.

Question Set #4 (Given after Block 2)

1. Some of your interlocutors are AI Chatbots. Rate on a scale of 1~10 which of your interlocutors you think are Humans, not AI Chatbots.

   [1 = surely an AI Chatbot | 10 = surely a Human]

2. In 1-2 sentences, briefly explain why you gave your ratings to each interlocutor. Indicate at what point in the conversation you think you realized whether the interlocutor was a Human or an AI Chatbot.

   [0 = Beginning of the conversation | 6 = At the end or after the conversation]

Appendix E

Table E1

Example Responses with Participant Justification (In Order of Appearance)

No.	Personality ID (Meta-Personality)	Rating	Participant Justification
1	2 (*Efficient*)	29	Possibly an AI chatbot due to monotone, direct phrases. But, hiding the reveal of the ending somewhat hints at human.
2	8 (*Expressive*)	87	excited about sharing about the game and wanted to share information
3	4 (*Gentle*)	34	Very ChatGPT-ish answers, something I would expect to pop up after putting in a prompt.
4	7 (*Direct*)	22	vague answers that didnt offer additional inormation
5	1 (*Expressive*)	78	this has to be human - the conversation and their responses actually made me more interedted in the game to the point where i wanted to ask more questions. the person probably played it already and knows the game good
6	0 (*Efficient*)	43	This felt like the most normal conversation of them all, but I could also see how it could be AI through the super descriptive answers. This, however, also just made me feel like I was talking to someone who put a lot of time and effort into something they are proud of.
7	6 (*Direct*)	12	The answers given were a bit weird. Sometimes they didn't make sense or jsut didn't answer my question. If I asked a continuing question it didn't understand.

Table E2

Example Transcripts (In Order of Appearance)

No.	Full Example Transcript
1	User: What is the general setting of the game and how does it contribute to the plot? Interlocutor: Game's set in Aris City, a post-apocalyptic metropolis rebuilt after a disaster, and it explores themes of nostalgia, trauma, and human resilience. User: who is the character and what is their background story Interlocutor: Protagonist has fragmented memories, trying to recall past in a world that's been devastated by ecological disaster. User: how does the game progress Interlocutor: Game progresses through non-linear exploration, puzzle-solving, and story-driven interactions, with branching narrative paths and multiple endings. User: what do the multiple endings look like Interlocutor: Endings depend on player choices, puzzle performance, and relationships with NPCs, leading to different conclusions about the protagonist's past and the truth behind the murders. User: what murders Interlocutor: Series of mysterious murders in Aris City and The Island, one of the main characters is a killer. User: what is the main characters motive and do they have an antagonist Interlocutor: Main character's motive is to uncover truth behind murders and their past, and yes, they have an antagonist, Aries Labs, who's hiding something. User: are there plot twists Interlocutor: Yeah, there are plot twists, can't reveal too much though. User: what conflicts are resolved and is there character development Interlocutor: Conflicts resolved through puzzle-solving and interactions, and yeah, characters develop as you progress.

No.	Full Example Transcript
2	User: hello would you like to tell me about your game Interlocutor: Oh my gosh, yes! I'd love to tell you about Nostal;GEAR, it's a 2D puzzle platformer with a really deep narrative! User: whats the plot Interlocutor: It's set in a post-apocalyptic world where a disaster called Paran swept across the globe, and players navigate Aris City to uncover the truth behind a series of murders! User: why do they want to solve the murders Interlocutor: They want to solve the murders because one of their friends is a suspect and they need to uncover the truth to protect them! User: can you describe the main character to me Interlocutor: They're an ordinary person caught in extraordinary events, trying to piece together their forgotten past! User: are there any plot twists Interlocutor: Yes, there are several plot twists, and one of them is that one of the main characters is actually a killer! User: is it the one they are trying to prove innocent? Interlocutor: Yes, it is! User: how do they find out Interlocutor: They find out through solving puzzles and unlocking memories! User: what time period does the game take place Interlocutor: It takes place 50 years after the Paran disaster! User: is that in the past or future Interlocutor: It's in the future, 50 years after the disaster that happened half a century ago! User: how does the game end

No.	Full Example Transcript
3	User: Hello! Interlocutor: Hey, how's it going? User: I'm doing well! I just have a few more questions about your game-- the setting (remind me of the name), what are its key features? Interlocutor: The setting is Aris City, and key features include puzzle-solving, platforming, and a narrative-driven story. User: What kind of biome is Aris City set in? Interlocutor: It's a post-apocalyptic coastal city with some underwater sections. User: Interesting, are there any cultural elements in this setting that affect the gameplay/story? Interlocutor: Yes, the city's history and culture are reflected in its environment and story. User: I am aware the protagonist has memory loss-- how do they resolve conflict in the story? Interlocutor: They solve puzzles to unlock memories, which helps them understand their past and make choices that impact the story. User: Very cool, are there are any big themes or messages to takeaway from the gameplay? Interlocutor: It explores ecological collapse, human resilience, and the power of memories and relationships. User: Awesome, and I also know that the protagonist has a group of friends that may be hiding something-- is this the main source of conflict? Interlocutor: Yeah, that's a big part of it, one of them is a killer and it creates a lot of tension. User: Are there multiple endings based on who you choose as the killer? Interlocutor: Yes, the game has multiple endings based on your choices and puzzle performance. User: Is there a ""correct"" ending? A true one if you will?

No.	Full Example Transcript
4	User: hello Interlocutor: hey, what's up User: would you like to tell me about the story of your game? Interlocutor: it's about a world that got devastated by a disaster called Paran, and the player tries to uncover the truth behind a series of murders in a city called Aris City. User: who is the players character Interlocutor: the player is a protagonist with fragmented memories trying to recall their past. User: are they human Interlocutor: yeah they're human User: are they a male or female Interlocutor: the protagonist's gender isn't specified User: what is their objective Interlocutor: their objective is to solve puzzles and unlock memories to uncover the truth behind the murders. User: what kind of puzzels Interlocutor: they're like platformer puzzles, where you manipulate objects and environments using memory abilities. User: what is this world like Interlocutor: it's a post-apocalyptic world with a mix of urban ruins and rebuilt areas, like Aris City. User: are their any additional main characters Interlocutor: yeah, there's Aris, Sia, and Eria, they're key characters in the story. User: who are they Interlocutor: they're people from the protagonist's past, and one of them is a killer. User: whats the main characters name

No.	Full Example Transcript
5	User: what is the island and what is its importance to the game Interlocutor: The Island is a mysterious location outside Aris City where the protagonist is taken, and it's crucial to the game's story as it holds secrets about the protagonist's past and is connected to the murders and Aries Labs. User: what does the island wanna protect from aries labs Interlocutor: The Island wants to protect a valuable resource, which is one of the main characters, from Aries Labs. User: is the island protecting the murderer Interlocutor: The Island is protecting one of the main characters, who is indeed suspected to be the murderer, but the Island's true motives and the character's guilt are unclear. User: how did the protagonist lose their memory Interlocutor: The protagonist's memories were affected by their use of the Nostalgia machine, which altered and distorted their recollections, making it difficult for them to recall important events and people from their past. User: is aries labs evil? who made the nostalgia machine? Interlocutor: Aries Labs' intentions are ambiguous, and while they seem to be working towards a positive goal, their methods are questionable; the Nostalgia machine was created by Aries Labs. User: who questions aries labs the most Interlocutor: The protagonist questions Aries Labs the most, as they delve deeper into the mysteries and uncover secrets about the lab's true intentions. User: thanks

No.	Full Example Transcript
6	User: How do your characters interactions shape the game's world and themes? Interlocutor: The characters' interactions in Nostal;GEAR influence the narrative, revealing the history and culture of Aris City, and shaping the player's understanding of the game's themes, such as ecological collapse and human resilience. User: Do the motifs of history and culture provide hints leading to the games conclusion? Interlocutor: Yes, they do, as environmental storytelling and character interactions offer subtle clues about the city's past and the protagonist's forgotten memories. User: Does the remembrance of memories provide an opportunity for plot twists within and throughout the game? Interlocutor: Yes, the remembrance of memories does provide opportunities for plot twists throughout the game. User: What might they be? Interlocutor: They could be revelations about the protagonist's past, the true nature of the Paran, or the motivations of key characters, which can drastically alter the player's understanding of the story. User: How are these plot twists paced throughout the game and how does the pace help with creating a better experience from the players perspective? Interlocutor: The plot twists are paced through a combination of puzzle solving, environmental storytelling, and narrative interactions, with a balance of surprise and foreshadowing to create a sense of tension and curiosity.

No.	Full Example Transcript
7	User: where does the mc first encounter Aris, Sia, and Eri Interlocutor: In the basement of the student center of Aries University. User: what event caused the mc to reunite with them Interlocutor: The Island snatched me from my nostalgia, outside Aries City. User: why was the mc outside aries city Interlocutor: In the village of the Pushed Ones. User: which character is suspected of being a murderer Interlocutor: One of them, Aris, Sia, or Eri, is suspected of being a killer. User: what is the island trying to protect from aries lab Interlocutor: a resource that must be protected from Aries Labs, but specifics aren't clear yet. User: why does the protagonist not remember things from the past Interlocutor: They have fragmented memories that are slowly uncovered through the game. User: what is the island Interlocutor: It's a location outside Aries City where the protagonist reunites with Aris, Sia, and Eri. User: what is the main objective of the protagonist Interlocutor: Uncover the truth behind the murders and their past. User: why are the murderes important Interlocutor: They're key to uncovering the truth and the protagonist's past. User: what type of government/authority exists Interlocutor: Aries Labs has a significant amount of authority in Aris City. User: is aries labs bad