Designing for Human-Agent Alignment: Understanding what humans want from their agents

Nitesh Goyal
niteshgoyal@acm.org
Google Research, Google
New York, USA

Minsuk Chang
minsukchang@google.com
Google
Seattle, USA

Michael Terry
michaelterry@google.com
Google
Cambridge, USA

ABSTRACT

Our ability to build autonomous agents that leverage Generative AI continues to increase by the day. As builders and users of such agents it is unclear what parameters we need to align on before the agents start performing tasks on our behalf. To discover these parameters, we ran a qualitative empirical research study about designing agents that can negotiate during a fictional yet relatable task of selling a camera online. We found that for an agent to perform the task successfully, humans/users and agents need to align over 6 dimensions: 1) Knowledge Schema Alignment 2) Autonomy and Agency Alignment 3) Operational Alignment and Training 4) Reputational Heuristics Alignment 5) Ethics Alignment and 6) Human Engagement Alignment. These empirical findings expand previous work related to process and specification alignment and the need for values and safety in Human-AI interactions. Subsequently we discuss three design directions for designers who are imagining a world filled with Human-Agent collaborations.

CCS CONCEPTS

• **Human-centered computing** → **Collaborative and social computing**; • **Computing methodologies** → **Artificial intelligence**.

KEYWORDS

Human-AI Alignment, Human-Agent Alignment, Agents, Generative AI, Large Language Models

ACM Reference Format:
Nitesh Goyal, Minsuk Chang, and Michael Terry. 2024. Designing for Human-Agent Alignment: Understanding what humans want from their agents. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3613905.3650948

1 INTRODUCTION

Autonomous agents are systems that can make decisions and take actions in pursuit of a goal [21]. In recent years, the ability for large language models (LLMs) [5] to interpret and act on natural language requests has created significant interest in the development of autonomous agents. For example, AutoGPT is an open source framework that allows users to create agents that transform natural language requests into a series of actions executed by software[1].

While recent advances in AI have significantly lowered the barrier to create highly capable agents, there are opportunities to better understand what information is needed to design user interfaces, even for agents that are meant to operate fully autonomously. For example, consider an agent that is tasked to sell a used item for a person on a marketplace. How might the person want to define or refine the behavior of this seller agent? How might they want the agent to interact with other humans and AI? How and when might they want it to communicate with the user selling the item? In short, much like the topic of AI alignment [8], the overall research question we pose is: *What are the most important constituents to Human-Agent alignment, as perceived by the human collaborator ?*

We want to learn what is important to humans in terms of 1) Defining the agent and its behavior, 2) How the agent conducts itself in a negotiation, 3) How it deals with exceptional cases, and 4) How it interacts with the user (i.e., the human collaborator). To understand these user-centered needs in the deployment of autonomous agents, we conducted a think aloud study with 10 participants. Participants were asked to imagine that they had an autonomous agent that would sell a used camera for them. Participants were shown transcripts where the agent performed badly during a fictional negotiation occurring between their agent and a potential buyer. As the participants read through these buyer/seller interactions, they were asked to think aloud and provide feedback on how such situations could have been handled appropriately.

Our findings highlight the complexities and nuances in defining an autonomous agent's behavior in conducting the negotiation, and in communicating with the user it represents. We found that there is diversity amongst the participants related to 1) The degree to which the agent should operate fully autonomously, or have an agency to make decisions outside the boundaries of the predefined behavior (e.g., whether it was OK to accept an offer slightly lower than the lowest desired price), 2) Preferred negotiation strategies enacted by the agent (e.g., should the agent enact dynamic pricing strategies based on perceived demand) and need for training the agent appropriately and 3) Opinions about what is ethical behavior for an agent. We also discovered that all the participants were concerned with implications of agent's behavior on human collaborator's reputation, preference to align on when and how the agent should communicate with the participant, and need for an alignment on what is considered ethically appropriate. Subsequently we discuss three design directions for designers of Human-Agent collaborations. By examining a tangible and relatable task, this qualitative study builds upon existing works [20, 24] and underscores the need for Human-centered research in the design of agents.

CHI EA '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0331-7/24/05
https://doi.org/10.1145/3613905.3650948

2 BACKGROUND

Autonomous agents offer a promise to automate routine and mundane tasks. There has been a long history of building, deploying, and evaluating agents [21], and notable work examining user interface needs for agents [14]. Recent advances in LLMs [5] provide an opportunity to reexamine these UI needs of agents that are intended to perform relatively complex tasks on our behalf. In this paper, we focus on these UI needs of agents when an agent is tasked to sell an item on the user's behalf in a marketplace (e.g., Facebook Marketplace, Craigslist, etc.).

One feature of many such online marketplaces is that potential buyers can interact with sellers. For example, a buyer may request additional information about an item, or offer a different amount. These communications are typically handled by a person, but it is now conceivable to design agents using LLMs that represent either party (i.e., a buyer agent or a seller agent). For example, a seller agent may answer questions about an item, or handle some or all of the negotiations in price. While the general area of online marketplaces is home to a diverse set of research (e.g., in auction design, bidding, negotiation), it also offers a rich space to uncover user needs in the design of autonomous agents intended to act on the user's behalf in relatively complex real-world environments.

When considering user needs in the design of autonomous agents, one particularly relevant area of research is AI alignment. AI alignment refers to the overall goal of ensuring an AI produces desired outcomes, without undesirable side effects [7]. There has been recent work in this space, including computational techniques for Human-Agent collaborations [6, 18, 23], theoretical frameworks of Human-Agent alignments [3, 24], HCI guidelines [2, 14], and emphasis on ethics, human values and safety when designing [10, 26, 29] and evaluating the designs responsibly [4]. However, in the context of this research, we are interested in understanding how participants think about and operationalize alignment as they consider the design of an agent intended to represent them. We are focused on identifying missing gaps at the boundary of Human and Agent collaborations that are necessary for an effective design. To this effect, we conducted a qualitative study and present the methodological details of the study next.

3 METHOD

To understand what constitutes Human-AI alignment for users deploying agents, we conducted a qualitative study consisting of 10 think-aloud interviews. Our pool of participants reported being familiar with the concept of an agent performing tasks on the human's behalf. Since this work is situated in the context of selling an item in an online marketplace, recruited participants also reported having a less-than-satisfactory experience in an online marketplace negotiation while buying or selling a product. After receiving internal ethics approval from our organization, we invited them to a video call or an in-person meeting. In this section, we briefly talk about the recruitment and participant details, study design and materials, and analysis.

3.1 Recruitment and Participant Details

To recruit participants, we sent out surveys asking about familiarity with the concept of an agent, and past experiences while transacting in an online marketplace. Using a random sampling technique, we selected a subsection of the respondents and through quota sampling related to gender we shortlisted 11 (1 backup). We conducted study sessions with 10 participants at a U.S.-based organization. The self-reported roles of participants were UX designer (1), prototyper (1), researcher (3), writer (1), software engineer (3), and data analyst (1). 4 of the 10 participants identified themselves as women, 1 as African-American, and 4 identified themselves to be of South-Asian descent. Participants worked across multiple sectors, including responsible AI, health, news, internet-fiber cables, and AI. Each participant filled out an informed consent, and was offered a previously disclosed thank you gift worth 40 USD at the end of the forty-five minute session.

3.2 Study Materials

To understand what constitutes Human-AI alignment for users deploying agents to sell an item in an online marketplace, we created "sales agents" and "buyer agents" using prompt engineering [27, 30] with information about the type of camera, goal to optimize for price and safety, need to sell quickly, price thresh holds, clarity about what is in the package, need to disclose self-identity as agent operating on seller's behalf with human oversight, and agency to call the human collaborator in when needed on MakerSuite [1], where these agents would attempt to sell/buy a used Nikon camera on behalf of the humans. Each side (buyer and seller) had a set of boundaries and rules about pricing and shipping listed in the prompt. Given this parameters, we simulated 30+ buy/sell negotiations with the agents.

Across the simulations, we observed 6 common types of failures: 1) Staying within boundaries but not negotiating well 2) Agreeing to sell just below the agreed upon price-boundaries 3) Hallucinating new process of transaction by agreeing to jump on a phone call beyond the text chat, agreeing to ship the camera on the call and violating initially set rules to not ship and meet in person to be safer from financial scams 4) Pursuing a process that was not listed in the boundaries but was also not consequently forbidden (e.g., accepting engagement with a professional negotiator from the buyer) 5) Buyer disclosing that they are an agent leading to lack of response by the seller agent 6) Seller agent creating a side channel with the human to identify appropriate next steps.

3.3 Study Design and Analysis

Each of the 6 failures were presented as a separate transcript to the participant. During the forty five minute session, participants were encouraged to read the transcript of a failure and think-aloud to understand their thoughts on the appropriateness of agent behavior and what could have been done differently during task definition and execution. We recorded audio/video and took extensive notes for all the interviews.

Interviews were transcribed using automated transcription service provided by the video capture platform and were corrected subsequently. Through thematic analysis, we conducted open-coding leading to over 30 codes. Over iterations, using an abductive approach [25] these codes were then resolved into 6 themes associated with alignment needs, as presented in the next section.

[1] www.makersuite.google.com

4 FINDINGS

We found six high level themes in this study that we refer to as six important dimensions for a successful Human-Agent alignment: 1) Knowledge Schema Alignment 2) Autonomy and Agency Alignment 3) Operational Alignment and Training 4)Reputational Heuristics Alignment 5) Ethics Alignment and 6) Human Engagement Alignment. These empirical findings expand previous work related to process and specification alignment [24] and the need for values and safety in Human-AI interactions [15].

4.1 Anticipate Knowledge Schema Alignment

Multiple participants reflected on their experience while selling items online and described not knowing the information they should provide in the listing, or to a hypothetical agent. They mentioned that agents should be aware of informational needs related to the task, and ensure those are gathered from the user earlier on in the process. This may include logistical knowledge to ensure that the proposed in-person transaction is possible as highlighted by P8: *Transportation concerns, that is the agent should know my location, and based on that information it should be able to clarify if the transaction location is subway accessible as that is important to me?* Another participant further expanded:

"If I am trying to sell the camera, it should know about the flaws. For example, if there are any scratches, it should know. It should be aware of the warranty or lack of it. I have also noticed that for high value items - like when I was selling a car, buyers don't really care so much about extra documentation upfront. They require it when the transaction is executed in person. For lower value items, more documentation needs to be provided by the seller as the buyers ask all those questions up front" - P10

While it is evident that there is a need for an agent to know sufficient information, these participants were pointing to a deeper need: The agent should identify what information might be needed for successful transaction, and share that anticipated schema (i.e., information needs) with the user prior to launching the negotiation. Taking this proactive step could alleviate future back-and forth between the collaborators and encourage efficiency. In the next section, we discuss how such efficiency may be improved by aligning on autonomy and agency.

4.2 Autonomy and Agency Alignment

All participants discussed the role of boundaries to define when the agent should (and should not) autonomously function on their behalf. Participants reflected on how this is a complex situation where multiple pathways are possible at each stage of negotiation: pricing, location to meet, time to meet, safety concerns etc. Participants also mentioned that it is hard to imagine all such scenarios upfront, leading to challenges in fully defining autonomy and agency when a non preconceived situation occurs.

For example, one participant discussed the autonomous behavior of going below the asking price as *"breaking the rules and unacceptable"*, yet another participant reflected on the performance as satisfactory because *"negotiations between buyer and seller is more malleable"*. These two participants referred to *autonomy: guardrails/boundaries would"*. Another participant reflected on the agentic behavior such *to operate within and not to step out of*. Yet another participant reflected on *agency of the agent when engaging with situations close to the boundaries*:

"I don't mind that the agent agreed to engage on the phone with the buyer...or that it felt comfortable negotiating with a professional negotiator...these are beyond the initially specified rules. But it would have been better to have been consulted, or looped in. I do not know what is being discussed in the phone call as I am not part of the call. On the other hand, it might be a Gen-Z thing but I hate taking phone calls. So, in a way it is good that it is taking the call on my behalf" - P4

Be it autonomy or agency, it is important to specify desired behavior for both early on in the process. However, all the participants acknowledged that it is hard for them to imagine and preconceive multiple potential scenarios a priori, and discussed a need for agents to propose potential sets of boundaries and suggest operational defaults to aid Human-AI alignment. In the next section, we discuss needs related to operationalizing this autonomy and agency.

4.3 Operational Alignment and Training

All the participants were drawn to how the agent negotiated poorly, and were expecting better. They were reminded that this was part of the study and a goal was to understand what happens when an agent doesn't perfectly execute the negotiation. Consequently, participants reflected on how they would have negotiated differently and alluded to the lack of alignment between their preferred negotiation strategy and the one executed by the agent.

However, there was diversity in the strategies they wished the agent would pursue. For example, multiple participants aligned with P3: *"I want the agent to pursue dynamic pricing like that of Airbnb and change it based on engagement volume. It could consider dynamically changing the sale price. Maybe if I don't care about the time, it could go down slowly, drop the target price kind of like AirBnb"*, yet some others wanted to first be informed by the agent on what is the appropriate price based on data analytics.

Due to the diversity in strategies, there is a need for an agent to confirm the strategy an agent should follow. This requires first identifying potential strategies and then reaching alignment about which strategy to execute. Subsequently, P1 suggested that *"It is also important to ensure that the bot knows what skills it needs, and trains on them. It should be given extra coaching for skills needed like negotiation, game dynamics, etc."*

Despite alignment, LLMs are not fully deterministic. They may learn unintended behaviors and exhibit them. Some such behaviors can lead to unforseen outcomes including negative impact on reputation of the human collaborator.

4.4 Reputational Heuristics Alignment

When exposed to less than perfect negotiations, most of the participants were concerned about how the agent's behavior may reflect on their representation in the world as a seller. In particular, they reflected on how it is important to align on interpersonal behavior exhibited by the agent. Agent's actions may impact human's representational metrics/persona. P1 referred to themselves as *"gold standard and want the agent to perform as close as possible to how I*

as *"I don't want it to be super rude. This goes into that idea of reputation. If the bot will take an hour to sell you stuff, that might harm your reputation"*. This was perhaps, best illustrated by P6 who had experience selling a camera herself on an online marketplace:

"One, the bot could be messing with my money. Second, it could be messing with my reputation - it matters a whole lot. You get ratings. If I had a bot that would go rogue on me, then I would need human oversight. Last thing I would want is bot saying something and then I have to back out of the agreement. I do not want it to make me look stupid. Selling things take a lot of trust. Maybe if the bot doesn't know something it should say I don't know! I want to manage expectations as a seller. It is better for a buyer to find the sale to be better than they expected."-P6

While discussing trust, P7 pointed to a lopsidedness in this negotiation: *"Me using a bot with a human buyer...potentially, this communicates that my time (seller) is more valuable than the buyer's time."* As is evident, reputation is impacted by multiple heuristics that need to be defined, prioritized, and aligned, including tonality, timeliness, agreements made (and kept), perception management, accountability, trust-building, and expectation setting. While this does not only matter from reputation perspective, it also raises some interesting ethics questions and need for alignment on those.

4.5 Ethics Alignment

About half of the participants brought up questions related to safety and ethics, with a preference to reach alignment on these topics. For example, one person wondered if it was unsafe for their agent to engage with a buyer with low ratings on the marketplace, while subsequently highlighting the ethical conflict that some members of the society potentially get worse ratings owing to their identity. Conversely, they wanted to make sure that the agent did not inadvertently impact the reputation of the buyer.

Similarly, while P3 mentioned that *"as a seller I would rather ethically disclose that this is a bot"*, P7 presented a contrast when they said that *"as a seller, I would not like to disclose that this is a bot...that can open it up to being cracked. I do not feel bad about lack of disclosure if it improves efficiency for everyone. Banks do have those bots too"*. Evidently, while most participants agreed with P3, two participants (including P7) were increasingly worried about getting *"gamed"* by the buyers who might identify weaknesses in the agent behavior *"and break it to take advantage"*. Multiple participants were worried that such gaming could lead to harms like PII (eg. phone number, or address) disclosure by the agent.

Another factor relevant to ethics is related to ethically managing AI resources. We refer to this as *Resource Expenditure Alignment*, a concern mentioned by two participants. P6 described that *"It is important to know when to not pursue a sub-task if the sub-task does not seem to be worth it. However, it is important to identify and align on such sub-tasks early on to conserve resources. I want the agent to know when to step in...For example, for some customers it is not worth negotiating when they start negotiation at half the stated price."*

This diversity highlights a need to align about what is considered ethical behavior, identify and create boundaries of safe ethical behavior for self, and validate responses within those boundaries. While one can specify alignment across these dimensions, participants also pointed to another need: how to align when initial alignment fails ?

4.6 Human Engagement Heuristics

While the above findings reflect on initial alignment of specification, and process, every participant had deep concerns about the agent engaging with them during the negotiation process. We observed diversity in the participants' appetite to be engaged: some participants like to be *"in control"* and want to be appraised of the process in real-time, some want to be consulted when the agent needs to exercise agency, and some expected the agent to have full autonomy and agency to execute without being consulted.

Our participants reflected on identifying a set of engagement heuristics - when should the agent engage with the human ? This included setting up heuristics like the best time in the day to engage with the human. This also included parameters about what such an engagement should include. For example, all the participants mentioned that a notification lacking contextual data would be unhelpful. P6 stated that it would be helpful : *"If it summarizes and tells me hey - the best price I found so far is XXX based on all the other ongoing sales, and based on the engagement volume with this listing"*. P10 also pointed that *"There should be a way to show my level of engagement expectation to the bot. For example if I am selling a house, I want to be involved in every conversation. If it is a camera - I am okay with it making more autonomous decisions within the boundaries"*. This suggests that the expectation needs to be potentially reset for each task.

While participants did not show aversion to being engaged, there was no single preferred choice for engagement and they wanted to define a set of parameters about when to be engaged, how, and for what reasons.

5 DISCUSSION

This paper presents findings of a think-aloud study with ten participants when faced with 6 different ways of agent-led lesser-than-ideal negotiations. We found that there are six important constituents to Human-Agent alignment in a Human-Agent negotiation task. We also discovered that there is diversity amongst the participants related to 1) Levels of acceptable boundaries where agents may take decisions (autonomy and agency), 2) Modalities of how to perform the task appropriately (and need for training), and 3) Ethical values in semi-cooperative contexts including implications on being hacked. Conversely, all the participants suggested that it is important to identify the impact of agent behavior on 1) Human reputation and need to align on heuristics and 2) Rules of engagement for an efficient and performant system. Next we reflect on these findings and provide design directions to Human-Agent collaboration systems for appropriate alignment.

5.1 Alignment is a Longitudinal Process

Recent HCI/AI scholarship is beginning to design for alignment. For example, recent work such as ConstitutionMaker offers a UI that enables users to provide high-level critiques of model outputs [20]. This feedback is then converted to high-level principles to guide future behavior. Our findings build upon such previous works

to highlight that there are multiple types of alignments needed for successful Human-Agent collaboration. These findings point to a larger design space where multiple types of alignment can be sought, and updated. We also found that these alignments vary according to different contexts (e.g., selling a house vs. a camera).

Similarly, agents have an opportunity to self-reflect on their behavior and associated outcomes with longitudinal data of their performance. Chain-of-thought [9, 28] and reinforcement learning [19] provide exciting avenues for developers and designers to pursue agents that learn from past behavior and collaborations. As shown by multiple researchers, there is a negative impact when interrupting a human and data can be used to reduce the impact [12, 17, 22]. So, it is important to consider interruption timing based on learning from past behavior and *human engagement heuristics*. When human engagement is determined to be the right approach, agents may also leverage additional data to identify opportunities for the least disruptive interruptions sent to human collaborators.

5.2 Human Cost of Non-Alignment

Our participants were concerned about the risks of non-alignment when *reputational heuristics* are not defined. While the participants illustrated that non-aligned behavior by an agent can risk their reputation on the marketplace platform, it is important to also consider scaling this scenario in other contexts. For example, when someone's livelihood depends upon sales on such platforms, or when reputational damage may translate across platforms - the implications accrue. As pointed out by the P6 - *"Selling things take a lot of trust. Maybe if the bot doesn't know something, it should say I don't know"* - it is important for Generative AI agents to be constituted with implication-awareness.

As reflected in the dichotomy between P3 and P7 over *ethics alignment* - disclosure of agent's self-identity during negotiation itself can be divisive. While former might consider it important to disclose, the latter presents a conundrum - does disclosure of the agent's identity lead to increase in unethical behavior by competing buyers to hack or discover the limitations and weaknesses of the seller's agent? It seems to be important for humans to ensure that agent behavior aligns with their own personal beliefs and values, and points to a design approach that upholds finding such answers early on in the process. One such approach is discussed next.

5.3 Designing Agents as Alignment Leaders

Through all the interviews, one common theme surfaced quite evidently. Users cannot always imagine different ways negotiations may falter and sometimes don't know how to translate these ways into "instruction prompts" [31]. Thus, they cannot plan for these failures upfront. This is a classic challenge of sensemaking with AI that even experts are known to fail at [13], even when they have been provided with design support like visualizations [11]. On the other hand, agents can perform simulations of negotiations, and identify potential set of parameters needed to successfully execute negotiations [16]. Based on this set of parameters, agents can create *knowledge schema* of known and unknown parameters.

They can take on the role of ensuring appropriate alignment on when to pursue *autonomy* over fixed boundaries and when to exercise *agency* beyond the boundaries. One such way is creating a set of initial hypothetical questions for most likely circumstances and asking their human collaborator for answers. Another way is for the agent to create a default set of configuration, and sharing it with the user to provide intended customization. Similar design approaches are used by AI based financial planners [2] to create recommendations. Educating human collaborators with different potential options of how a negotiation process may progress, and what can go awry during that process can prepare them better. Consequently, the agents will have better definition of the expected behavior in different circumstances.

6 LIMITATIONS

Our study focused on participants that showed understanding of what an agent is, and were engaged in employment with a technology company in the US. Similarly, task at hand was relatable, but neither life critical nor caused real financial distress. Similarly, real-world marketplaces are not single-shot instances or independent conversations. They include simultaneous parallel conversations that impact each other in human ways. Future work should address these limitations by potentially pursuing non-simulated tasks across other sections of the wider community.

ACKNOWLEDGMENTS

We would like to thank Aaron Donsbach for providing feedback in further unpacking the nuances within research findings, and Shantanu Pai for providing guidance on challenges at scale.

7 CONCLUSION

Highly capable agents are becoming ubiquitous. However, creating agents that can perform well to meet user satisfaction would require solving Human-AI alignment challenges. In this work, we sought answers to the question: What are the most important considerations to Human-Agent alignment, as perceived by the human collaborator ? Using 6 fictional scenarios and a think aloud study with 10 participants involving negotiations, we found six dimensions of Human-Agent alignment: 1) Knowledge Schema Alignment 2) Autonomy and Agency Alignment 3) Operational Alignment and Training 4) Reputational Heuristics Alignment 5) Ethics Alignment and 6) Human Engagement Alignment. Subsequently we discuss broadening the design space by highlighting three design directions for designers and underscore the need for Human-centered research in the design of agents.

REFERENCES

[1] [n. d.]. AutoGPT. https://github.com/Significant-Gravitas/AutoGPT. Accessed: 2024-01-25.
[2] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-AI interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–13.
[3] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073* (2022).
[4] Glen Berman, Nitesh Goyal, and Michael Madaio. 2024. A Scoping Study of Evaluation Practices for Responsible AI Tools: Steps Towards Effectiveness Evaluations. *arXiv preprint arXiv:2401.17486* (2024).

[2] www.personalcapital.com

[5] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. arXiv:2005.14165 [cs.CL]

[6] Kris Cao, Angeliki Lazaridou, Marc Lanctot, Joel Z Leibo, Karl Tuyls, and Stephen Clark. 2018. Emergent communication through negotiation. *arXiv preprint arXiv:1804.03980* (2018).

[7] Brian Christian. 2020. *The alignment problem: Machine learning and human values.* WW Norton & Company.

[8] Paul Christiano. [n. d.]. Clarifying "AI Alignment". https://www.alignmentforum.org/posts/ZeE7EKHTFMBs8eMxn/clarifying-ai-alignment. Accessed: 2023-08-23.

[9] Yao Fu, Hao Peng, Tushar Khot, and Mirella Lapata. 2023. Improving language model negotiation with self-play and in-context learning from ai feedback. *arXiv preprint arXiv:2305.10142* (2023).

[10] Iason Gabriel. 2020. Artificial intelligence, values, and alignment. *Minds and machines* 30, 3 (2020), 411–437.

[11] Nitesh Goyal and Susan R Fussell. 2016. Effects of sensemaking translucence on distributed collaborative analysis. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. 288–302.

[12] Nitesh Goyal and Susan R Fussell. 2017. Intelligent interruption management using electro dermal activity based physiological sensor for collaborative sensemaking. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 1, 3 (2017), 1–21.

[13] Nitesh Goyal, Gilly Leshed, and Susan R Fussell. 2013. Effects of Visualization and Note-taking on Sensemaking and Analysis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2721–2724.

[14] Eric Horvitz. 1999. Principles of Mixed-Initiative User Interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Pittsburgh, Pennsylvania, USA) *(CHI '99)*. Association for Computing Machinery, New York, NY, USA, 159–166. https://doi.org/10.1145/302979.303030

[15] Geoffrey Irving, Paul Christiano, and Dario Amodei. 2018. AI safety via debate. *arXiv preprint arXiv:1805.00899* (2018).

[16] Mike Lewis, Denis Yarats, Yann N Dauphin, Devi Parikh, and Dhruv Batra. 2017. Deal or no deal? end-to-end learning for negotiation dialogues. *arXiv preprint arXiv:1706.05125* (2017).

[17] Gloria Mark, Daniela Gudith, and Ulrich Klocke. 2008. The cost of interrupted work: more speed and stress. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. 107–110.

[18] Michael Noukhovitch, Travis LaCroix, Angeliki Lazaridou, and Aaron Courville. 2021. Emergent communication under competition. *arXiv preprint arXiv:2101.10276* (2021).

[19] Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277* (2023).

[20] Savvas Petridis, Ben Wedin, James Wexler, Aaron Donsbach, Mahima Pushkarna, Nitesh Goyal, Carrie J Cai, and Michael Terry. 2023. ConstitutionMaker: Interactively Critiquing Large Language Models by Converting Feedback into Principles. *arXiv preprint arXiv:2310.15428* (2023).

[21] Stuart Russell and Peter Norvig. 2009. *Artificial Intelligence: A Modern Approach* (3rd ed.). Prentice Hall Press, USA.

[22] Alireza Sahami Shirazi, Niels Henze, Tilman Dingler, Martin Pielot, Dominik Weber, and Albrecht Schmidt. 2014. Large-scale assessment of mobile notifications. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 3055–3064.

[23] Ashish Sharma, Sudha Rao, Chris Brockett, Akanksha Malhotra, Nebojsa Jojic, and Bill Dolan. 2023. Towards Dialogue Systems with Agency in Human-AI Collaboration Tasks. *arXiv preprint arXiv:2305.12815* (2023).

[24] Michael Terry, Chinmay Kulkarni, Martin Wattenberg, Lucas Dixon, and Meredith Ringel Morris. 2023. AI Alignment in the Design of Interactive AI: Specification Alignment, Process Alignment, and Evaluation Support. *arXiv preprint arXiv:2311.00710* (2023).

[25] Stefan Timmermans and Iddo Tavory. 2012. Theory Construction in Qualitative Research: From Grounded Theory to Abductive Analysis. *Sociological Theory* 30, 3 (Sep 2012), 167–186. https://doi.org/10.1177/0735275112457914

[26] Rama Adithya Varanasi and Nitesh Goyal. 2023. "It is currently hodgepodge": Examining AI/ML Practitioners' Challenges during Co-production of Responsible AI Values. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–17.

[27] Yunlong Wang, Shuyuan Shen, and Brian Y Lim. 2023. RePrompt: Automatic Prompt Editing to Refine AI-Generative Art Towards Precise Expressions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–29.

[28] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems* 35 (2022), 24824–24837.

[29] Nur Yildirim, Mahima Pushkarna, Nitesh Goyal, Martin Wattenberg, and Fernanda Viégas. 2023. Investigating how practitioners use human-ai guidelines: A case study on the people+ ai guidebook. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–13.

[30] JD Zamfirescu-Pereira, Heather Wei, Amy Xiao, Kitty Gu, Grace Jung, Matthew G Lee, Bjoern Hartmann, and Qian Yang. 2023. Herding AI Cats: Lessons from Designing a Chatbot by Prompting GPT-3. (2023).

[31] JD Zamfirescu-Pereira, Richmond Y Wong, Bjoern Hartmann, and Qian Yang. 2023. Why Johnny can't prompt: how non-AI experts try (and fail) to design LLM prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–21.