Unpacking AI Alignment

Matt Bilik
University of Washington
bilik@uw.edu

Pinar Barlas
Western University

Reuben Binns
University of Oxford

Emily Tseng
University of Washington
emtseng@uw.edu

Alexandra Olteanu
Mila - Quebec AI Institute
alexandra@aolteanu.com

Abstract

AI alignment has become a central focus of AI research and practice—cast as a response to the unprecedented scale, perceived autonomy, and generality of AI systems. *The impulses of AI alignment work are, however, not new*. The very act of developing computational systems involves an intention about *what* these systems should or should *not* do (e.g., be helpful, not be harmful), which entails some notion of system *behavior specification*. Making sure computational systems do what we want them to do has thus long been a core goal of computing research. Regardless of the form it takes, effective behavior specification requires an understanding of *why* certain specification choices are needed, and facilitates the verification and validation of a given implementation to ensure the system behaves as intended.

While AI alignment functions as a behavior specification device, alignment work often underspecifies or leaves implicit the very components that define the task of alignment. To help researchers and practitioners interrogate current approaches to AI alignment and understand what they do, we present an explanatory framework that makes explicit the prototypical components integral to *any* alignment project—such as what is being aligned, along which system property, according to what criteria—alongside the definitional (*what*), operational (*how*), and epistemic (*why*) choices made to instantiate or implement those components. By doing so, the framework surfaces possible conceptual, operational, and epistemic missteps involved in these choices, which can inform debates about these choices and, more broadly, about whether alignment is an effective behavior specification device for AI systems.

1 AI alignment as a behavior specification device for AI systems

Alignment has become a central focus of AI research and practice—and a topic of intense debate.[1] This is evidenced by the surge in AI alignment research across fields [e.g., 4, 10, 21, 27, 42, 49, 56], alongside a growing number of industry, governmental, and media reports declaring that AI alignment is critical to the design, development, and deployment of AI systems [e.g., 9, 14, 24].[2] AI alignment work is often cast as a response to the unprecedented scale, perceived autonomy, and generality of AI systems—traits that are taken to demand new and different solutions.

[1] In AI research and practice, the term *alignment* has been used to loosely refer to work that ranges from steering AI systems towards some intended goals—i.e., focused on how to produce desired system behavior [e.g., 6, 7, 25, 41]—to determining whether the behavior of AI systems is aligned to those goals—i.e., focused on how to verify that an AI system actually produces that desired behavior [e.g., 3, 30, 47, 54].

[2] For simplicity, we use the term AI system even when the AI system consists of only a single AI model.

Preprint.

We argue that the impulses of AI alignment work are not new. The very act of designing and developing computational systems necessarily involves an intention about *what* these systems should do—such as generating personalized, helpful, or culturally appropriate outputs—or should *not* do—such as generating harmful, misleading, or inaccurate outputs. Making sure computational systems do what we want them to do—which necessarily involves some notion of system *behavior specification*—has thus long been a core goal of computing research. Why, then, is AI alignment different from the ways in which the computing community has traditionally approached system behavior specification?[3] What seems *new* about AI alignment are the assumptions and beliefs that animate it, such as the view that AI systems should, or *necessarily* will, have human-like qualities or capabilities [8, 12, 15, 22, 35, 39, 53] and, as a result, that AI systems require practitioners to specify appropriate system behaviors along complex criteria such as human values, culture, or emotions [e.g., 25, 32, 50, 54].

Regardless of the form the system behavior specification takes, its main role is to *define what the system should do* [2, 18, 20, 33, 37, 44, 55, 57]—such as via precise descriptions of *what* its behavior should be. This, in turn, facilitates the validation of *how* the specification is implemented in practice (i.e., to either produce or evaluate whether the system behavior adheres to that specification). Still, effective behavior specification is difficult without a good understanding of *why* certain specification choices are needed given a system's intended use [44, 52]. Furthermore, the specification of complex systems—which current AI systems often are—also typically relies on multiple components, such as functional and non-functional requirements (e.g., behaviors the system must perform versus what performance the system should maintain), or acceptance criteria (e.g., the condition for the system behavior to be considered appropriate) [e.g., 18, 20, 23, 55].

However, despite AI alignment's position as the dominant specification device for determining the behavior of AI systems, it is often unclear what constitutes an aligned AI system and, thus, how a practitioner should build one. This is, in no small part, due to AI alignment work often leaving *underspecified* or *implicit* its key constitutive components (see §2). Understanding whether AI alignment—as is currently conceptualized and operationalized—*is* or *can* be made effective at ensuring AI systems behave as we want them to, demands clarity about the theoretical, epistemic, and methodological commitments that underpin AI alignment work.

2 Explanatory framework for AI alignment as a behavior specification device

The distinction between the description of *what* the behavior of a system should be (specification) from *how* the system is implemented to produce the desired behavior (implementation), is reminiscent of the distinction measurement theory makes between constructs (or *what* is being measured) and their operationalization (*how* they are measured) [1, 13, 26]. To identify the constitutive components of AI alignment work—as it pertains to both producing and assessing *aligned* behaviors—and tease apart the definitional (*what*), operational (*how*), and epistemic (*why*) choices made in selecting and implementing these components, we drew on system behavior specification literature from software engineering [e.g., 18, 44, 55] and on the measurement theory literature from the social sciences [e.g., 1, 26, 36, 45].

With these distinctions in mind, we examined a purposive sample of AI alignment research[4] to identify and separate (1) any choices researchers made about *what* aligning an AI system is intended to do—and thus any choices about the conceptualization of the underlying constructs that those instantiations of alignment relied on—from (2) choices related to *how* alignment work operationalizes desired system behaviors. This examination also foregrounded questions about *why* or what informed and motivated any choices the researchers made. By open coding and thematically clustering these choices, we found that—regardless of the method used to produce or assess aligned system behaviors—AI alignment work appears to (implicitly or explicitly) rely on several key constitutive components (highlighted with *emphasized text* in the paragraph below). We organized these components and the relationships they share in an explanatory-analytical framework (illustrated in Figure 1):

[3] We use the term *behavior specification* in its broadest sense, without committing researchers and practitioners to a certain approach to or type of system behavior specification or operationalization.

[4] We identified a purposive sample of 19 AI alignment-focused papers [i.e., 3, 6, 7, 11, 17, 19, 25, 28–32, 34, 38, 42, 47, 48, 50, 51] that either (i) gained visibility in relevant disciplinary venues (e.g., NeurIPS, ACL, FAccT), or (ii) include authors known for their work on AI alignment, or (iii) discuss substantially different criteria or approaches to alignment (e.g., socio-affective vs. cultural alignment, or determining alignment by relying on notions of internal consistency vs. by comparing with human actions), or (iv) match a mix of any of these criteria.

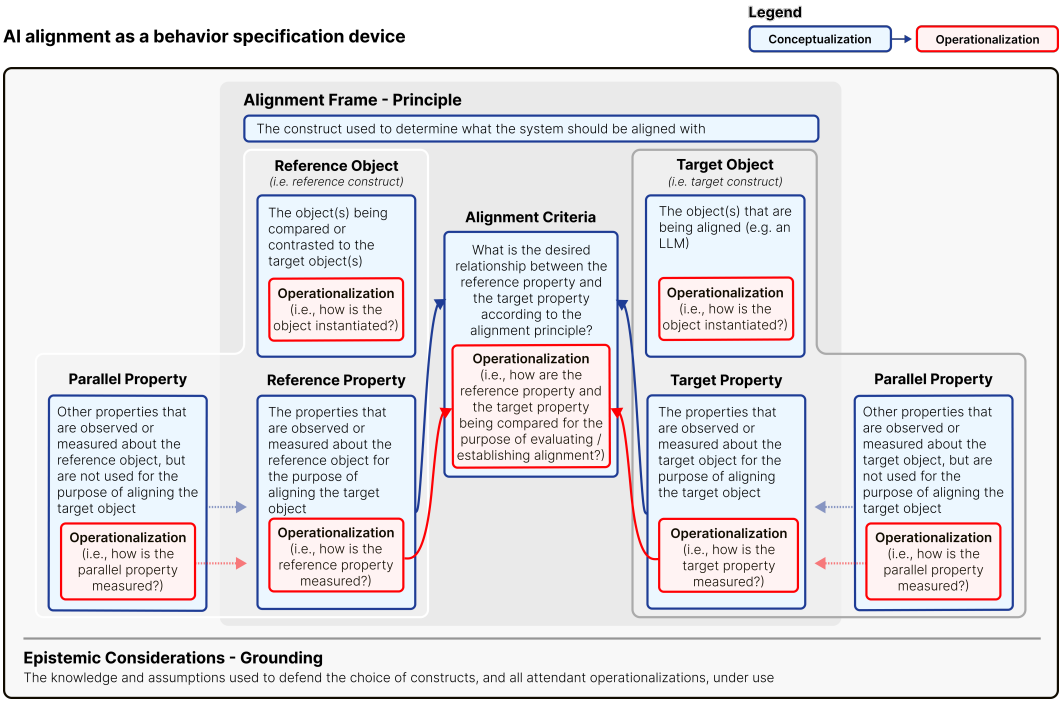


Figure 1: Explanatory-analytical framework for examining and describing AI alignment work through the lenses of system behavior specification and measurement theory. It foregrounds and makes explicit key constitutive components that AI alignment work commonly relies on to specify what the desired behavior should be, whose behavior is being specified—or who or what is being aligned—and how to assess whether they are aligned. Each gray box describes the role of each constitutive component at the conceptual level, while each corresponding red box describes the instantiation or operationalization of that component. The framework also illustrates relationships between components through both arrows and nested boxes; relationships that need to be preserved not only at the conceptual level, but also at the operational level. Both the conceptual and operational choices related to these constitutive components are informed by a diversity of epistemic considerations.

AI alignment work commonly describes the desired behavior for an AI system in relation to the behaviors or characteristics of humans or other AI systems, and according to an underlying *alignment principle* such as human values, preferences, or culture. Specifically, the alignment of an AI system involves a comparison between a *target* (often an AI model) and a *reference object*[5](often a human or a group of humans).[6] This comparison is often done by assessing the degree to which the properties of these objects—which we refer to as *target* and *reference properties*—are in some way analogous to each other, typically by determining whether their relationship satisfies a certain set of *alignment criteria*. For instance, the *alignment criteria* may require an AI system to generate the same answers to a national values survey—where the answers operationalize the *target* and *reference properties*—as the respondents in a given country did [3, 34, 50]. Sometimes the goal of aligning an AI system may also be constrained by other objectives, such as the preservation of other *parallel properties*, like language or math understanding [e.g., 25].

3 Examining and unpacking AI alignment work

Our framework (Figure 1) provides researchers, practitioners, and other stakeholders (such as users, policymakers, or journalists) the analytical scaffolding to examine and scrutinize the conceptual and operational choices made in AI alignment work, and the assumptions those choices rely on. It

[5]We use the term *object* to refer to who or what is aligned, or compared to. It can be a model, a human, or a group of humans.

[6]The *reference object* is often left implicit with only the *reference property* being described. For example, a *reference property* described as "beliefs about what is harmful" implies a group of humans (the *reference object*) who hold those beliefs [11].

Missteps	Construct-level	Nomological-level
Conceptual (what)	***Related to whether the constructs under use are known, clearly defined, and reasonable.*** Missteps here can result from different sources of conceptual ambiguity such as whether constructs 1) are clearly articulated (*lack of conceptual clarity*), 2) are underspecified or too broad (*lack of conceptual specificity*), 3) change meaning as work—or the article describing the work—progresses (*lack of conceptual consistency*), 4) are distinguishable from other concepts (*lack of conceptual separability*), or 5) are incompatible or not-fit-for-purpose (*conceptual collisions*).	***Related to whether constructs are conceptually related to each other, and whether that relation is known, clearly defined, and reasonable.*** Missteps here can result from different sources of nomological ambiguity such as 1) unknown or unspecified relations between constructs (*nomological underspecification*), 2) ascription of certain properties to objects that cannot have those properties e.g., green ideas, emotional LLMs (*category errors*), 3) assuming constructs are related in ways they cannot be related, including by definition (*nomological incompatibilities*), or 4) drawing conclusions that are unsupported by the relations between constructs (*false analogies*).
Operational (how)	***Related to whether the constructs' operationalization is appropriately described, justified, and validated.*** Missteps here can result from different sources of content validity concerns such as whether 1) the operationalization attempts to capture a construct whose meaning is known, stable, and uncontested (*contestedness and operationalism*), 2) construct validity was attempted or established (*lack of construct validation*), 3) the operationalization captures additional properties not related to the construct (*construct contamination*), or 4) the operationalization fails to capture relevant aspects of the construct (*construct underrepresentation*).	***Related to whether the constructs' operationalization appropriately preserves the theoretical relations between them.*** Missteps here can result from different sources of nomological validity concerns such as whether 1) both the constructs and the theoretical relations between them are known, stable, and uncontested (*conceptual and nomological ambiguities*), or 2) the constructs' operationalization preserves both the theoretical relations between different aspects of a composite construct (*structural validity*) and 3) the way constructs are intended to relate with each other (*nomological validity*).
Epistemic (why)	***Related to whether the knowledge, beliefs, and assumptions—including about how constructs under use relate to each other—that inform conceptual and operational choices are known and reliable.*** Missteps here can result from different sources of epistemological concerns such as 1) reliance on the belief that humans and their characteristics can be computationally represented via fixed properties (*essentialization*), 2) treatment of abstract constructs as though they have concrete computational representations (*reification*), 3) reliance on assumptions that are known to be theoretically implausible (*epistemic contamination*), 4) reliance on practices and assumptions that lack empirical evidence or formal validation even if they are widely accepted (*folk epistemology* and lack of *epistemic dependability*), or 5) underspecification of the knowledge and assumptions the work relies on (*epistemic opaqueness*).	

Table 1: AI alignment missteps matrix, highlighting possible *conceptual* (or definitional), *operational*, and *epistemic* missteps related to choices researchers and practitioners can make about any of the constitutive components involved in producisng or assessing aligned behaviors, at both construct and nomological-level. For each category of missteps, the table foregrounds common sources of concern.

does so (1) by making explicit the prototypical, constitutive components involved in AI alignment work—e.g., what is being aligned (or, as we term it, the *target object*), along which system property (the *target property*), and according to what *alignment criteria*—and (2) by clarifying the definitional, operational, and epistemic choices that practitioners make to create each of these components. In doing so, our framework also helps surface possible missteps related to common conceptual, operational, and epistemic choices.

To demonstrate the utility of our framework, we organized possible missteps—which we observed while constructing the framework (§2) and re-applying it to the purposive sample of research papers on alignment—into a misstep matrix (Table 1). This matrix distinguishes not only whether these concerns have to do with conceptual, operational, or epistemic choices, but also whether they are related to the constructs themselves (*construct-level*) or to the relationships that the constructs are assumed to share (*nomological-level*).

Conceptual missteps are often related to the degree to which the constructs that alignment work relies on are known, clearly defined, and reasonable. For example, competing understandings of what value-aligned behavior might be can make it unclear whether such behavior entails 1) taking actions that reflect a set of values, 2) not taking actions that violate the values (e.g., via refusals), 3) negating any actions that would violate the values, 4) or taking actions that neither reflect nor violate the values [e.g., 11, 16, 48]. *Operational* missteps relate to whether the operationalization of the constructs is appropriately justified and validated. For example, culture and cultural values—which have been conceptualized in a diversity of ways [e.g., 3, 5, 17, 29, 34, 50]—are commonly used as the alignment principle and/or the target and reference properties. However, construct validity concerns arise when we only know how these constructs are operationalized but not how they are defined [e.g., 30], or when several different conceptualizations are concomitantly under use [e.g., 3], making it unclear which specific conceptualization the operationalization is attempting to capture. *Epistemic* missteps relate to whether the knowledge and assumptions that motivate the work are known and reliable. Missteps here include, for example, epistemic contamination or folk epistemological bias, such as when conceptual choices rely on contested, unexamined, or theoretically implausible assumptions. A recurrent theme is treating AI systems as having the ability to posses human characteristics like a capacity for "*thoughts*" and "*underlying mindsets*" [54] despite little evidence or compelling reasons to think that existing AI systems actually have them [40, 43, 46].

4 Conclusions

The impulses of AI alignment work are not new. Because AI alignment work is intended to ensure or assess whether AI systems behave as intended, *AI alignment functions as a behavior specification*

device. We presented a framework that helps illustrate the key constitutive components—such as the alignment principle, the reference and alignment objects and their respective properties, and the alignment criteria—that AI alignment work relies on, alongside the conceptual, operational, and epistemic choices made in selecting and implementing those components.

By illustrating these constitutive components alongside the choices the practitioners and researchers make, our framework helps foreground a series of possible missteps in AI alignment work around what is being specified (*conceptual*), how it is specified and implemented (*operational*), and why (*epistemic*). These missteps highlight a need for the field (a) to establish standards for AI alignment research and practice that help prevent these missteps as the field continues to expand and, more broadly, (b) to investigate whether AI alignment—as it is currently construed—*is* or *can* be made effective at ensuring AI systems behave as we want them to.

References

[1] Robert Adcock and David Collier. Measurement validity: A shared standard for qualitative and quantitative research. *American political science review*, 95(3):529–546, 2001.

[2] Ahmed Ahmed, Kevin Klyman, Yi Zeng, Sanmi Koyejo, and Percy Liang. Speceval: Evaluating model adherence to behavior specifications. *arXiv preprint arXiv:2509.02464*, 2025. URL https://arxiv.org/abs/2509.02464.

[3] Badr AlKhamissi, Muhammad ElNokrashy, Mai Alkhamissi, and Mona Diab. Investigating Cultural Alignment of Large Language Models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12404–12422, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.671. URL https://aclanthology.org/2024.acl-long.671/.

[4] Usman Anwar, Abulhair Saparov, Javier Rando, Daniel Paleka, Miles Turpin, Peter Hase, Ekdeep Singh Lubana, Erik Jenner, Stephen Casper, Oliver Sourbut, et al. Foundational challenges in assuring alignment and safety of large language models. *arXiv preprint arXiv:2404.09932*, 2024. URL https://arxiv.org/abs/2404.09932.

[5] Arnav Arora, Lucie-aimée Kaffee, and Isabelle Augenstein. Probing Pre-Trained Language Models for Cross-Cultural Differences in Values. In Sunipa Dev, Vinodkumar Prabhakaran, David Ifeoluwa Adelani, Dirk Hovy, and Luciana Benotti, editors, *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 114–130, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.c3nlp-1.12. URL https://aclanthology.org/2023.c3nlp-1.12/.

[6] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback, April 2022. URL http://arxiv.org/abs/2204.05862. arXiv:2204.05862 [cs].

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

[8] Yoshua Bengio. The Consciousness Prior, December 2019. URL http://arxiv.org/abs/1709.08568. arXiv:1709.08568 [cs.LG].

[9] Yoshua Bengio, Stephen Clare, Carina Prunkl, Maksym Andriushchenko, Ben Bucknall, Malcolm Murray, Rishi Bommasani, Stephen Casper, Tom Davidson, Raymond Douglas, et al. International ai safety report 2026. *arXiv preprint arXiv:2602.21012*, 2026. URL https://arxiv.org/abs/2602.21012.

[10] Leonard Bereska and Efstratios Gavves. Mechanistic interpretability for ai safety–a review. *arXiv preprint arXiv:2404.14082*, 2024. URL https://arxiv.org/abs/2404.14082.

[11] Jan Betley, Daniel Tan, Niels Warncke, Anna Sztyber-Betley, Xuchan Bao, Martín Soto, Nathan Labenz, and Owain Evans. Emergent Misalignment: Narrow finetuning can produce broadly misaligned LLMs. *Nature*, 649(8097):584–589, January 2026. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-025-09937-5. URL http://arxiv.org/abs/2502.17424. arXiv:2502.17424 [cs].

[12] Borhane Blili-Hamelin, Christopher Graziul, Leif Hancox-Li, Hananel Hazan, El-Mahdi El-Mhamdi, Avijit Ghosh, Katherine A. Heller, Jacob Metcalf, Fabricio Murai, Eryk Salvaggio, Andrew J. Smart, Todd Snider, Mariame Tighanimine, Talia Ringer, Margaret Mitchell, and Shiri Dori-Hacohen. Position: Stop treating 'AGI' as the north-star goal of AI research. In *Proceedings of the 42nd International Conference on Machine Learning*, pages 81090–81117. PMLR, October 2025. URL https://proceedings.mlr.press/v267/blili-hamelin25a.html.

[13] Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. Stereotyping norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1004–1015, 2021.

[14] Samuel R Bowman, Megha Srivastava, Jon Kutasov, Rowan Wang, Trenton Bricken, Benjamin Wright, Ethan Perez, and Nicholas Carlini. Findings from a pilot anthropic—openai alignment evaluation exercise. *Anthropic Alignment Science Blog*, 2025.

[15] Erik Brynjolfsson. The turing trap: The promise & peril of human-like artificial intelligence. In *Augmented education in the global age*, pages 103–116. Routledge, 2023.

[16] Bochuan Cao, Yuanpu Cao, Lu Lin, and Jinghui Chen. Defending against alignment-breaking attacks via robustly aligned LLM, 2024. URL https://openreview.net/forum?id=V01FPV3SNY.

[17] Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. Assessing Cross-Cultural Alignment between ChatGPT and Human Societies: An Empirical Study. In Sunipa Dev, Vinodkumar Prabhakaran, David Ifeoluwa Adelani, Dirk Hovy, and Luciana Benotti, editors, *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 53–67, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.c3nlp-1.7. URL https://aclanthology.org/2023.c3nlp-1.7/.

[18] Marsha Chechik, Eunsuk Kang, Shahar Maoz, Jan Oliver Ringert, and Allison Sullivan. Specification engineering: Foundations for the future of software development (dagstuhl seminar 25392). *Dagstuhl Reports*, 15(9):160–182, 2026.

[19] Nicholas Clark, Hua Shen, Bill Howe, and Tanushree Mitra. Epistemic Alignment: A Mediating Framework for User-LLM Knowledge Delivery, April 2025. URL http://arxiv.org/abs/2504.01205. arXiv:2504.01205 [cs].

[20] Edmund M. Clarke and Jeannette M. Wing. Formal methods: state of the art and future directions. *ACM Comput. Surv.*, 28(4):626–643, December 1996. ISSN 0360-0300. doi: 10.1145/242223.242257. URL https://doi.org/10.1145/242223.242257.

[21] Juntao Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. Safe rlhf: Safe reinforcement learning from human feedback. In *International Conference on Learning Representations*, volume 2024, pages 50750–50777, 2024.

[22] Chaz Firestone. Performance vs. competence in human–machine comparisons. *Proceedings of the National Academy of Sciences*, 117(43):26562–26571, 2020.

[23] Martin Glinz. On non-functional requirements. In *15th IEEE international requirements engineering conference (RE 2007)*, pages 21–26. IEEE, 2007.

[24] Ryan Greenblatt, Carson Denison, Benjamin Wright, Fabien Roger, Monte MacDiarmid, Sam Marks, Johannes Treutlein, Tim Belonax, Jack Chen, David Duvenaud, et al. Alignment faking in large language models. *arXiv preprint arXiv:2412.14093*, 2(1), 2024. URL https://arxiv.org/abs/2412.14093.

[25] Saffron Huang, Divya Siddarth, Liane Lovitt, Thomas I. Liao, Esin Durmus, Alex Tamkin, and Deep Ganguli. Collective Constitutional AI: Aligning a Language Model with Public Input. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, pages 1395–1417, New York, NY, USA, June 2024. Association for Computing Machinery. ISBN 979-8-4007-0450-5. doi: 10.1145/3630106.3658979. URL https://dl.acm.org/doi/10.1145/3630106.3658979.

[26] Abigail Z. Jacobs and Hanna Wallach. Measurement and Fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 375–385, Virtual Event Canada, March 2021. ACM. ISBN 978-1-4503-8309-7. doi: 10.1145/3442188.3445901. URL https://dl.acm.org/doi/10.1145/3442188.3445901.

[27] Jiaming Ji, Tianyi Qiu, Boyuan Chen, Jiayi Zhou, Borong Zhang, Donghai Hong, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Lukas Vierling, Zhaowei Zhang, Fanzhi Zeng, Juntao Dai, Xuehai Pan, Hua Xu, Aidan O'Gara, Kwan Ng, Brian Tse, Jie Fu, Stephen Mcaleer, Yanfeng Wang, Mingchuan Yang, Yunhuai Liu, Yizhou Wang, Song-Chun Zhu, Yike Guo, Yaodong Yang, and Wen Gao. Ai alignment: A contemporary survey. *ACM Comput. Surv.*, 58(5), November 2025. ISSN 0360-0300. doi: 10.1145/3770749. URL https://doi.org/10.1145/3770749.

[28] Zachary Kenton, Tom Everitt, Laura Weidinger, Iason Gabriel, Vladimir Mikulik, and Geoffrey Irving. Alignment of Language Agents, March 2021. URL http://arxiv.org/abs/2103.14659. arXiv:2103.14659 [cs].

[29] Aidan Kierans, Avijit Ghosh, Hananel Hazan, and Shiri Dori-Hacohen. Quantifying Misalignment Between Agents: Towards a Sociotechnical Understanding of Alignment. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(26):27365–27373, April 2025. ISSN 2374-3468. doi: 10.1609/aaai.v39i26.34947. URL https://ojs.aaai.org/index.php/AAAI/article/view/34947.

[30] Jiseon Kim, Jea Kwon, Luiz Felipe Vecchietti, Alice Oh, and Meeyoung Cha. Exploring Persona-dependent LLM Alignment for the Moral Machine Experiment, April 2025. URL http://arxiv.org/abs/2504.10886. arXiv:2504.10886 [cs].

[31] Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, Bertie Vidgen, and Scott A. Hale. The PRISM Alignment Dataset: What Participatory, Representative and Individualised Human Feedback Reveals About the Subjective and Multicultural Alignment of Large Language Models, December 2024. URL http://arxiv.org/abs/2404.16019. arXiv:2404.16019 [cs].

[32] Hannah Rose Kirk, Iason Gabriel, Chris Summerfield, Bertie Vidgen, and Scott A. Hale. Why human-AI relationships need socioaffective alignment, February 2025. URL http://arxiv.org/abs/2502.02528. arXiv:2502.02528 [cs].

[33] Percy Liang. On the elusiveness of a specification for AI. Talk, NIPS 2015 Symposium: Algorithms Among Us, 2015. URL https://www.microsoft.com/en-us/research/video/symposium-algorithms-among-us-percy-liang/.

[34] Reem I. Masoud, Ziquan Liu, Martin Ferianc, Philip Treleaven, and Miguel Rodrigues. Cultural Alignment in Large Language Models: An Explanatory Analysis Based on Hofstede's Cultural Dimensions, May 2024. URL http://arxiv.org/abs/2309.12342. arXiv:2309.12342 [cs].

[35] Qiaozhu Mei, Yutong Xie, Walter Yuan, and Matthew O Jackson. A turing test of whether ai chatbots are behaviorally similar to humans. *Proceedings of the National Academy of Sciences*, 121(9):e2313925121, 2024.

[36] Samuel Messick. Validity. *ETS Research Report Series*, 1987(2):i–208, 1987.

[37] B. Meyer. On Formalism in Specifications. *IEEE Software*, 2(1):6–26, January 1985. ISSN 0740-7459. doi: 10.1109/MS.1985.229776. URL http://ieeexplore.ieee.org/document/1695257/.

[38] Piotr Mirowski, Juliette Love, Kory Mathewson, and Shakir Mohamed. A Robot Walks into a Bar: Can Language Models Serve as Creativity SupportTools for Comedy? An Evaluation of LLMs' Humour Alignment with Comedians. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, pages 1622–1636, New York, NY, USA, June 2024. Association for Computing Machinery. ISBN 979-8-4007-0450-5. doi: 10.1145/3630106.3658993. URL https://dl.acm.org/doi/10.1145/3630106.3658993.

[39] Nils J Nilsson. Human-level artificial intelligence? be serious! *AI magazine*, 26(4):68–75, 2005.

[40] Alexandra Olteanu, Su Lin Blodgett, Agathe Balayn, Angelina Wang, Fernando Diaz, Flavio Calmon, Margaret Mitchell, Michael Ekstrand, Reuben Binns, and Solon Barocas. Rigor in ai: Doing rigorous ai work requires a broader, responsible ai-informed conception of rigor. *Advances in Neural Information Processing Systems*, 38, 2026.

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

[42] Xiangyu Qi, Ashwinee Panda, Kaifeng Lyu, Xiao Ma, Subhrajit Roy, Ahmad Beirami, Prateek Mittal, and Peter Henderson. Safety alignment should be made more than just a few tokens deep. In *International Conference on Learning Representations*, volume 2025, pages 54911–54941, 2025.

[43] Inioluwa Deborah Raji, I Elizabeth Kumar, Aaron Horowitz, and Andrew Selbst. The fallacy of ai functionality. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, pages 959–972, 2022.

[44] Alessandra Russo. An introduction on the role of formal specifications in software engineering. Lecture notes, Software Engineering – Design I (Course 220), Imperial College London, 2011/2012. Available at http://www.doc.ic.ac.uk/~ar3/lectures/Sed/L1/IntroductionPrintOut.pdf.

[45] Olawale Salaudeen, Anka Reuel, Ahmed Ahmed, Suhana Bedi, Zachary Robertson, Sudharsan Sundar, Ben Domingue, Angelina Wang, and Sanmi Koyejo. Measurement to Meaning: A Validity-Centered Framework for AI Evaluation, June 2025. URL http://arxiv.org/abs/2505.10573. arXiv:2505.10573 [cs.CY].

[46] Murray Shanahan. Talking about large language models. *Communications of the ACM*, 67(2): 68–79, 2024.

[47] Hua Shen, Nicholas Clark, and Tanushree Mitra. Mind the Value-Action Gap: Do LLMs Act in Alignment with Their Values?, September 2025. URL http://arxiv.org/abs/2501.15463. arXiv:2501.15463 [cs].

[48] Hua Shen, Tiffany Knearem, Reshmi Ghosh, Yu-Ju Yang, Nicholas Clark, Tanushree Mitra, and Yun Huang. ValueCompass: A Framework for Measuring Contextual Value Alignment Between Human and LLMs, November 2025. URL http://arxiv.org/abs/2409.09586. arXiv:2409.09586 [cs].

[49] Tianhao Shen, Renren Jin, Yufei Huang, Chuang Liu, Weilong Dong, Zishan Guo, Xinwei Wu, Yan Liu, and Deyi Xiong. Large language model alignment: A survey. *arXiv preprint arXiv:2309.15025*, 2023. URL https://arxiv.org/abs/2309.15025.

[50] Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. Cultural bias and cultural alignment of large language models. *PNAS Nexus*, 3(9):pgae346, September 2024. ISSN 2752-6542. doi: 10.1093/pnasnexus/pgae346. URL https://doi.org/10.1093/pnasnexus/pgae346.

[51] Michael Terry, Chinmay Kulkarni, Martin Wattenberg, Lucas Dixon, and Meredith Ringel Morris. Interactive AI Alignment: Specification, Process, and Evaluation Alignment, September 2024. URL http://arxiv.org/abs/2311.00710. arXiv:2311.00710 [cs].

[52] Axel Van Lamsweerde et al. Goal-oriented requirements engineering: A guided tour. *re*, 1:249, 2001.

[53] Iris van Rooij, Olivia Guest, Federico Adolfi, Ronald de Haan, Antonina Kolokolova, and Patricia Rich. Reclaiming AI as a Theoretical Tool for Cognitive Science. *Computational Brain & Behavior*, 7(4):616–636, December 2024. ISSN 2522-087X. doi: 10.1007/s42113-024-00217-5.

[54] Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, Jiangjie Chen, Cheng Li, and Yanghua Xiao. InCharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1840–1873, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.102. URL https://aclanthology.org/2024.acl-long.102/.

[55] Karl Eugene Wiegers and Joy Beatty. *Software requirements*, volume 2. Microsoft press Redmond, 2013.

[56] Yotam Wolf, Noam Wies, Oshri Avnery, Yoav Levine, and Amnon Shashua. Fundamental limitations of alignment in large language models. *arXiv preprint arXiv:2304.11082*, 2023. URL https://arxiv.org/abs/2304.11082.

[57] Pamela Zave and Michael Jackson. Four dark corners of requirements engineering. *ACM transactions on Software Engineering and Methodology (TOSEM)*, 6(1):1–30, 1997.