Decision

Dimensions of Disagreement: Divergence and Misalignment in Cognitive Science and Artificial Intelligence

Kerem Oktar, Ilia Sucholutsky, Tania Lombrozo, and Thomas L. Griffiths

Online First Publication, August 15, 2024. https://dx.doi.org/10.1037/dec0000244

AMERICAN PSYCHOLOGICAL ASSOCIATION

Decision

ISSN: 2325-9965

https://doi.org/10.1037/dec0000244

Dimensions of Disagreement: Divergence and Misalignment in Cognitive Science and Artificial Intelligence

Kerem Oktar[1], Ilia Sucholutsky[2], Tania Lombrozo[1], and Thomas L. Griffiths[1, 2]

[1] Department of Psychology, Princeton University

[2] Department of Computer Science, Princeton University

Our understanding of disagreement is rooted in psychological studies of human behavior, which typically cast disagreement as divergence: two agents forming diverging *evaluations* of the same object. Recent work in artificial intelligence highlights how disagreement can also arise from misalignment in how agents *represent* that object. Here, we formally describe these two dimensions of disagreement, clarify the relationship between them, and argue that strategies for conflict resolution and collaboration are likely to be ineffective (or even backfire) if they do not consider misalignment in representations. Moreover, we identify how taking misalignment into account can enrich current research on judgment and decision making, from biased advice taking to algorithm aversion, and discuss implications for artificial intelligence research.

Keywords: disagreement, divergence, misalignment, representation

What is disagreement? It is intuitive to think of disagreement as a divergence of judgment: If Deniz believes that vaccines are safe and Sade does not, then they disagree. This intuitive notion of divergence undergirds much work on disagreement in judgment and decision-making (JDM) research (e.g., Reeder et al., 2005), as well as political science, social psychology, and epistemology (Carothers & O'Donohue, 2019; Frances & Matheson, 2019; Iyengar et al., 2019). Here, we argue that developments in artificial intelligence (AI) and computational cognitive science highlight another dimension of disagreement—representational misalignment—that formalizes ideas rooted in philosophy and developmental psychology (e.g., Carey, 1985; Kuhn, 1962), and with important implications for conflict resolution and collaboration. We begin our discussion with a historical case study to illustrate these two notions of disagreement and explain why the distinction matters.

Divergence Versus Misalignment

In 1663, Galileo was convicted of heresy by the Roman Catholic Inquisition for his belief that the sun is the center of the universe, as Pope Urban VIII (the voice of God on Earth) instead maintained that the center of the universe is the Earth (heliocentrism vs. geocentrism; see Finocchiaro, 1989). Galileo (G) and Pope Urban (U) clearly disagreed. Through a Bayesian lens, whereby beliefs are conceptualized as subjective probability assignments, we can characterize the extent of this disagreement through divergences in their credences about whether the sun is the center of the universe (S; e.g., divergence = $|P_G(S) - P_U(S)|$; see Frances & Matheson, 2019).

Divergence parsimoniously captures the way disagreement has been conceptualized and operationalized in much psychological research, from

Kerem Oktar https://orcid.org/0000-0002-0118-5065
Ilia Sucholutsky https://orcid.org/0000-0003-4121-7479
Tania Lombrozo https://orcid.org/0000-0001-5637-1431
Thomas L. Griffiths https://orcid.org/0000-0002-5138-7255

The authors have no known conflicts of interest to disclose.

Kerem Oktar played a lead role in conceptualization, investigation, visualization, writing–original draft, and writing–review and editing. Ilia Sucholutsky played a supporting role in conceptualization, visualization, writing–original draft, and writing–review and editing. Tania Lombrozo played a supporting role in visualization and writing–review and editing. Thomas L. Griffiths played a supporting role in visualization and writing–review and editing.

Correspondence concerning this article should be addressed to Kerem Oktar, Department of Psychology, Princeton University, Princeton, NJ 08540, United States. Email: oktar@princeton.edu

disagreement over policy preferences (Reeder et al., 2005) to statistical estimates (Minson et al., 2011) and aesthetic judgments (Cheek et al., 2021), among others. Across these cases, the literature typically treats proximate, convergent judgments as "agreement," and distant, divergent judgments as "disagreement."

Yet, divergence fails to capture discrepancies in the algorithms and representations underlying people's judgments. For instance, early heliocentric and geocentric models of the universe differed greatly in the astronomical structures they implied, despite making highly *convergent* predictions about the apparent positions of planets in the solar system *relative to the Earth* (Gearhart, 1985). Consequently, if we assessed whether Galileo and Pope Urban disagree about the solar system by measuring their divergence on astronomical predictions made from Earth (such as whether there will be a solar eclipse on a particular date), we would reach the conclusion that they agree, as the probabilities they assign to most events would be very similar. Galileo and Pope Urban could have differing explanations for why they believe what they believe, based on different representations or procedures, yet converge. Convergence can thus mask deep disagreements rooted in misaligned representations, and fail to fully characterize disagreement judgments (Oktar et al., 2024). We next formalize these dimensions of disagreement.

Computing Disagreement

Divergence can be formalized as a distance metric on beliefs about the world (e.g., Euclidean distance; for other distance metrics, see Deza & Deza, 2009). It is typically measured through distances in individual judgments. Minson et al. (2011), for instance, operationalize disagreement by computing the quantitative differences in a dyad's estimates (e.g., about the average income of Israeli families).

Misalignment can be formalized as a dissimilarity measure between representations across agents, or the same agent over time (for a review, see Sucholutsky et al., 2023). It is typically measured through correlations of pairwise similarity judgments in a circumscribed task or domain, with items rated as more similar interpreted as being closer to each other in representational space—a method spearheaded by Shepard (1980) and currently implemented through representational similarity analysis (Kriegeskorte et al., 2008). Brandt (2022), for instance, operationalized misalignment in politics by estimating pairwise associations across political concepts (e.g., gay rights and gun ownership) and conducting representational similarity analysis across people's associations. Misalignment can be assessed at different scales—for instance, focusing on a narrow domain (e.g., representation of the solar system) or a broader one (e.g., representation of the Milky Way), and evaluated with respect to coarse-grained stimuli (e.g., similarity of planets) or fine-grained stimuli (e.g., similarity of planet composition, trajectories, habitability).

The Relationship Between Divergence and Misalignment

Divergence and misalignment capture disagreement through measures of distance and dissimilarity (where proximate judgments correspond to convergence, and similar representations correspond to alignment). The relationship between the two depends on how expansively we measure misalignment. To illustrate this point, consider an important controversy: vaccine laws.

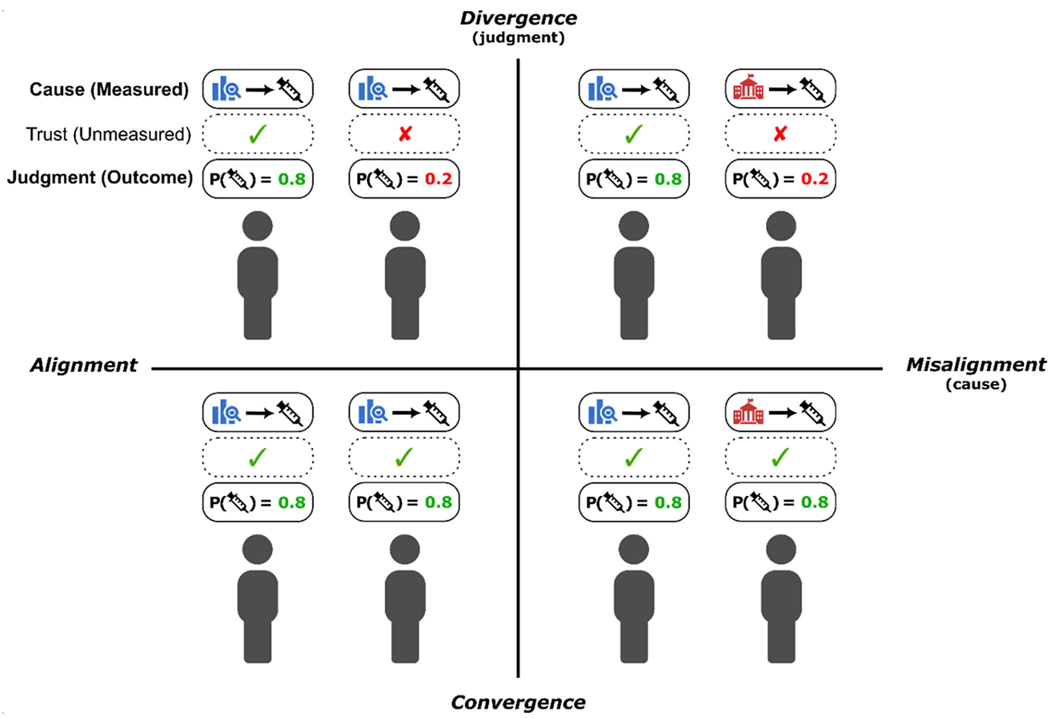
Should vaccines be mandatory? An individual's representation of this topic could include causal models of vaccine development (e.g., whether they are the result of scientific research or manufactured for profit; Loomba et al., 2021), intuitive beliefs about diseases (e.g., how dangerous they are; Powell et al., 2023), moral commitments and values (e.g., about the importance of autonomy; Akande et al., 2022), among others (see Fasce et al., 2023). Practically, our evaluations of misalignment cover only a subset of these components. Divergence can therefore occur despite alignment in *measured* components due to differences in *unmeasured* components or their processing.[1]

[1] At the theoretical limit, evaluations of alignment capture all measurable components and essentially compare the entire, expressible, judgment-relevant mental states of two agents. At this limit of maximal coverage, alignment implies convergence if the agents reason similarly (since the relevant mental states are practically identical), but misalignment does not imply divergence (as in our astronomical example). On the other hand, at the limit of minimal coverage, dissimilarity is measured with respect to a single component—if we pick the component to be the judgment itself, misalignment and divergence would be identical; if we pick it to be some other component, the discussion of practical evaluation above applies. Relatedly, minimizing divergence does not necessarily imply complete alignment, as there may be unmeasurable components of representations (e.g., phenomenological experiences; see Figure 2). We bracket these philosophical questions here; see Meißner (2023) and Poldrack (2021) for further discussion.

For example, if we measured misalignment concerning vaccine laws by focusing on beliefs about vaccine development, we might find that two people are aligned if both believe that scientific research is responsible for vaccine development. Yet, their judgments on vaccine mandates could diverge if they have differing levels of trust in the reliability of academic research—an aspect of their representations that was not measured. This is a common occurrence in everyday conflicts, where we know we disagree, think we understand why, but in fact fail to appreciate the latent nuances underlying others' perspectives (Epley & Caruso, 2008). Similarly, misalignment can occur despite convergence: Two people could have different causal representations of which institutions are responsible for vaccine development, but both could trust the relevant institutions, such that they generate the same judgment concerning whether vaccines should be mandated (see Figure 1).

The preceding discussion clarifies when divergence and misalignment can come apart, but an important question remains: What is the direction of the relationship? Intuitively, differences in judgments follow from one's relevant mental representations, so divergence should follow from misalignment. Though this will hold *synchronically*, there are feedback loops that complicate the causal picture, such that bidirectional relationships can arise *diachronically*. For example, people use their judgments to make inferences about their preferences (rationalization; see Cushman, 2020). Similarly, people may use divergence itself to update or create representations that can in turn support future judgments. For example, imagine a debate between a Republican and a Democrat on vaccine mandates. The Republican may diverge

Figure 1
Misalignment and Divergence Over Vaccine Mandates



Note. Figure shows misalignment and divergence in the case of vaccine attitudes. Divergence is shown as discrepancies in judgments (differences in probability assignments, with 0.8 indicating belief in the statement and 0.2 disbelief), and measured misalignment is shown as discrepancies in beliefs about the causal processes generating vaccine research (either scientific research, in blue, or corporate profits, in red). Note that the case of divergence despite alignment arises from differences in unmeasured components (trust in institutions; shown in italics and dashed ellipses). See the online article for the color version of this figure.

from the Democrat's position mid debate and later make inferences about their own representations based on this divergence (e.g., inferring that they value bodily autonomy). Divergence and misalignment thus have a bidirectional causal relation over time, though misalignment causes divergence at the point a judgment is made.

Why Distinguishing Divergence and Misalignment Matters

The scope of representational misalignment extends far beyond planets and vaccinations. From variation in our experience of basic sensations to our understanding of abstract concepts, diversity in the structure of mental representations has challenged theories of knowledge, communication, and ethics for millennia (Meißner, 2023; Poldrack, 2021). Developmental, educational, and cross-cultural psychologists have faced the daunting task of characterizing intuitive theories and mechanisms of change when these theories are not only distinct but potentially incommensurable (e.g., Carey, 1991; Vosniadou et al., 2008). Yet, some misalignments are more practically impactful than others. In the case of sensation, for instance, there can be variation in the phenomenal experience of the same stimuli (e.g., what seems red to me may seem green to you; Zaman et al., 2021). This variation is unproblematic from the standpoint of communication and collaboration if the relative structure of internal representations is preserved (e.g., we both agree that red is more similar to purple than to green; Goldstone & Rogosky, 2002; see Figure 2).[2] In the following sections, we therefore focus on cases where misalignment has important consequences for: (a) the design of conflict-resolving interventions, (b) our explanations of important phenomena in JDM, or (c) the engineering of artificial agents that collaborate with humans.

Designing Conflict-Resolving Interventions

Recognizing misalignment can clarify whether, when, and how disagreements can be resolved. From a divergence-focused perspective, for instance, a straightforward approach to reducing disagreement is providing disagreeing agents with a common set of data that relate to the issue in question. This strategy for disagreement reduction is known as the "deficit model" in science communication (Simis et al., 2016; see also Farrell et al., 2019; Hartman et al., 2022). The intuitive appeal of this strategy is so strong that early work in social psychology took *greater* disagreement in response to the same set of data to defy "any normative strategy imaginable for incorporating new evidence relevant to one's beliefs" (Ross & Anderson, 1982, p. 145).

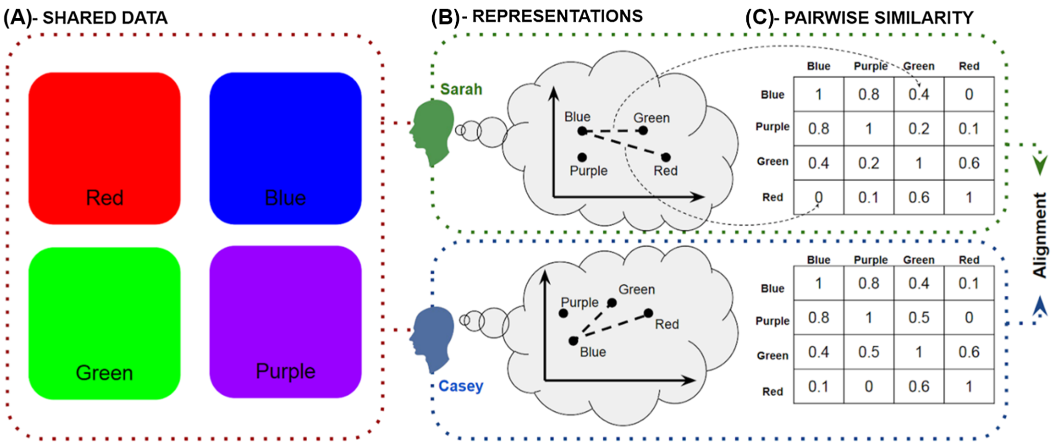
Yet, such polarization can be both common and rational in the presence of misalignment. Differences in the set of alternative hypotheses being represented or the conditional dependencies between hypotheses and data being considered can lead to divergent conclusions from the same data (Jaynes, 2003, Ch. 5.2; Jern et al., 2014). For instance, consider two people with different prior beliefs about whether a source is likely to be reliable, with one assuming that the testimony of the source tends to be positively correlated with truth and the other assuming that it is negatively correlated. Upon hearing the source argue for anthropogenic climate change, their credences will move in opposing directions, with more data leading to greater polarization (for a related finding, see Cook & Lewandowsky, 2016; the same pattern can also arise from inferences of bias, see Oktar et al., 2024). The source of polarization in this case is different representations of the relationship between the source's testimony and truth.

In the case of astronomy, a similar mechanism led scholars to entrenched, persistent disagreement despite observing the same stellar data. Whereas geocentric models represented stars as being relatively close to earth, heliocentric models took them to be very far. Thus, the observation that stars remain the same size year-round confirmed both geocentrists' views (if the earth is in the center, stars should be the same size as they are always equidistant), as well as heliocentrists' views (if the stars are very far, they will appear to be the same size since the orbit of the earth is too small to make an observable difference; see Grant, 1984).

Whether observations will push us toward polarization, entrenchment, or agreement thus depends on both our representations and the kind of data we observe. Scientists are intimately

[2] We can formally frame this mismatch in the following way. If we consider each stimulus as being internally encoded as a vector (e.g., of neural activations), variation in sensation would correspond to differences in the absolute values of these vectors: For instance, your red vector may be equivalent to my green vector. What matters for practical alignment (e.g., communication and collaboration) is whether the relative distances between vectors is preserved across the two representational spaces.

Figure 2
Computing Representational Alignment From Pairwise Similarity



Note. Representational alignment as a measure of differences in pairwise similarity judgments. Though Sarah and Casey perceive individual colors differently, their representations are structurally similar, as captured by the correspondence in their pairwise similarity matrices (except they are minorly misaligned in their representations of purple)—most measures of misalignment that are commonly used are sensitive to such structural similarity. Adapted from "Alignment With Human Representations Supports Robust Few-Shot Learning," I. Sucholutsky and T. L. Griffiths, 2023. arXiv preprint (https://doi.org/10.48550/arXiv.2301.11990). Copyright by The Authors. See the online article for the color version of this figure.

familiar with this fact: Some experimental data efficiently discriminate between hypotheses; many others do not (Platt, 1964; cf. O'Donohue & Buchanan, 2001). As described above, observations of most stars were not diagnostic with respect to heliocentrism versus geocentrism—but some data points could speak to one theory over the other (a prediction regarding the cycles of Venus; Gingerich, 2011). Similarly, interventions for promoting mutual understanding need to provide data that allow people to discriminate between competing representations. For example, providing further scientific data about the benefits of vaccines may not resolve a disagreement between an antivaxxer who believes that science is corrupt and a scientifically inclined family member—but providing evidence that science is a relatively unbiased process may be more effective (e.g., Ranney & Clark, 2016; cf. Gershman, 2019).

How can we know which data will be most impactful? Artificial intelligence research has developed methods for generating stimuli that maximize divergence between AI systems and humans (e.g., Goodfellow et al., 2014), and recent work provides Bayesian methods for generating stimuli that maximally differentiate representations across agents in simple domains (e.g., face perception; see Golan et al., 2022). Generalizing such approaches to discovering parts of semantic space that lie at the heart of misalignment is an important direction for future research.

The more general point that complex representations are underdetermined by simple evaluations poses a challenge for theories of JDM quite broadly (Richters, 2021). Tests of foundational theories (Tversky & Kahneman, 1974) typically manipulate simple stimuli (e.g., the probabilities in a gamble between two options; Peterson et al., 2021), which result in the elicitation of simple representations that are relatively consistent across individuals. Ongoing work aims to extend classical theories to more complex choice tasks using representations elicited from large language models (LLMs), and in so doing, capture informative individual differences in decision making (e.g., Bhatia, 2023). Future research could generalize these advances to develop better models of judgment and belief as well.

Enriching Extant Research: Implications for Advice Taking and Algorithm Aversion

As mentioned above, much research in JDM takes a divergence-first approach to disagreement. Yet, taking misalignment into account can enrich current lines of inquiry while raising novel questions about the nature of disagreement. For example, research on advice taking has investigated how people weigh their own judgments versus those of others in estimation tasks (Bonaccio & Dalal, 2006), as well as those of human versus algorithmic advisors (Glikson & Woolley, 2020). People typically overweight their own judgments in these studies, and many mechanisms have been suggested to account for this egocentric bias, from asymmetric access to reasons (Yaniv & Kleinberger, 2000) to biased sampling (Hütter & Ache, 2016) and motivated reasoning (Kappes et al., 2020). Misalignment offers a distinct and synergistic explanation for biased advice taking.

When receiving advice, people jointly learn *from* and *about* their advisors (Bovens & Hartmann, 2003). For example, if an advisor provides contradictory advice about similar problems, we might simultaneously use their advice and grow suspicious of their reliability (see Orchinik et al., 2023). Beyond merely estimating reliability, we might also use their estimates to infer their representation of the problem space—in a simple estimation task with one predictor and one outcome, for instance, we could sample the advisor's estimates across possible values of the predictor to infer their representation of the function relating the two variables. To illustrate, imagine trying to allocate loans to customers based on a credit score, *C*, and advice from an unfamiliar advisor, Logan. To evaluate whether you should trust Logan, you could see how his recommendations fluctuate with *C*—if Logan's recommendations closely track *C*, that would tell you that he has a representation of the problem that perhaps aligns with yours (e.g., in assuming that high credit scores indicate responsible financial habits). If Logan's recommendations correlate negatively with *C*, however, you might infer that his representation is misaligned. This could lead you to discount Logan's advice, especially if it leads to a divergent judgment that is incorrect.

The literature on algorithm aversion shows exactly this pattern: Algorithms are not discounted until they make unexpected and atypical mistakes, after which people quickly lose confidence in them (Dietvorst et al., 2015; cf. Logg et al., 2019). Beyond divergence, inferences of misalignment could thus contribute to understanding how people utilize others' advice. With the advent of AI assistants powered by LLMs, there is now also a rapidly growing literature exploring how trust in an AI affects whether people use it in decision making (Choudhury & Shamszare, 2023) and the relationship between the accessibility of representations and trust (Zou et al., 2023).

Raising Novel Questions

Reducing disagreement to divergence simplifies inferences of and from disagreement—and incorporating misalignment raises questions by complicating this analysis. Whereas divergence can be approximated with one sample or communicative act, misalignment is much more difficult to estimate.[3] Minimally, it requires observing systematic divergence across a range of judgments. Maximally, it entails inferring or even fully simulating the other agent's internal representation of the task. This naturally raises the question of if, when, and how people go through this more informationally and computationally intensive inference process, rather than using divergence-based heuristics.

An important factor may be the ease of generalization from one's own internal representation to that of the other agent. Generalizing to similar agents in well-known domains may be the easiest case since one's own representations can be leveraged to estimate alignment. Intuitively, I may be able to put myself in my best friend's shoes when discussing an issue we are both familiar with, but understanding how the Fair Isaac Corporation credit scoring algorithm represents credit-worthiness may be much more difficult. This is because in the former case, I can use my own representations as a basis for inferring those of my friend (Goldman, 2006; Woo & Mitchell, 2020), whereas in the latter case, I do not have the requisite knowledge or mechanisms for understanding artificial agents in unfamiliar domains. Relatedly, the extent of representational misalignment for word meanings predicts failures of communication across people (Duan & Lupyan, 2023).

[3] Relatedly, teachers are fairly accurate at tracking what students know and do not know but are much less accurate at recognizing their alternative understandings and models (Chi et al., 2004). For example, many children represent the circulatory system as comprised of simply the heart and the body, without a special role for the lungs in providing oxygenation to blood. Teachers are better at detecting factual inaccuracies (e.g., that oxygenated blood flows from atria to ventricles) than they are at diagnosing the presence of flawed representations (e.g., models where the lungs are not a part of the circulatory system).

As for inferences from disagreement, misalignment raises new questions about the striking tendency for individuals to persist in their beliefs amid dissent (e.g., roughly 90% do not question their views upon contemplating societal disagreement; Oktar & Lombrozo, 2022b). A common path to persistence is subjectivity: If I believe that euthanasia is morally permissible and that moral beliefs are matters of subjective opinion, I may persist in my views despite disagreement. But what grounds such inferences? We are not aware of formal treatments that explain what judgments of subjectivity track (except for recent research in philosophy, see Sytsma et al., 2021). One possibility is that subjectivity tracks irreconcilable representational diversity—in domains where there is a lot of variance in how people perceive issues or stimuli (e.g., on abstract notions like morality or love), and where there is no basis for evaluating which representations are more accurate or practically useful, people may expect disagreement to be incommensurable (for research on diversity in human concepts, see Marti et al., 2023). Domains with representational uniformity and tools for adjudicating better or worse representations (e.g., formally defined systems like games, financial markets, or mathematics) may prove more conducive to conciliation. Beyond disagreement resolution, understanding the roots of subjectivity would have widespread implications for our understanding of JDM quite broadly, from how people make decisions across domains (Oktar & Lombrozo, 2022a) to how they evaluate moral beliefs (Goodwin & Darley, 2012).

Implications of the Distinction for AI Research

With increasingly powerful and inscrutable AIs being deployed in real-world settings, there is mounting concern about the risks of relying on these systems for decision making. Accordingly, the question of "value alignment" has received much attention in recent academic AI research (e.g., Bommasani et al., 2021; Gabriel, 2020), and has become a focus in industry (with OpenAI recently investing 20% of its compute on a new "superalignment" team; Sutskever & Leike, 2023). Despite the name, most recent work on value alignment instead focuses on preventing *divergence*. For instance, LLMs owe much of their success to the use of reinforcement learning with human feedback (RLHF), whereby the unsupervised output of these models is constrained by human evaluations of model outputs in the fine-tuning stage (Ouyang et al., 2022). Note, however, that this technique ultimately corresponds to divergence reduction: The model is trained to prioritize outputs that are close to the judgments of the humans providing feedback. This raises the worry that RLHF may "render models aligned 'on the surface,' and that they still harbor harmful biases or other tendencies that may surface in more subtle contexts" (Bai et al., 2022, p. 35). In other words, we could end up developing models that are radically misaligned and diminish our capacity for detecting such misalignment due to training procedures that disincentivize the expression of divergence from human judgment.[4]

For instance, RLHF trains Generative Pretrained Transformers to explicitly denounce racist, sexist, and biased rhetoric (Fang et al., 2023), but recent research has shown that these models nevertheless retain biased latent associations in their representations. Turpin et al. (2023) constructed pairs of ambiguous stories where one of two suspicious characters was responsible for a crime, and the only difference across the stories was that the race and gender of the characters were flipped. When asked to identify which character was guilty, LLMs consistently picked the stereotypically targeted group (e.g., Black men) versus alternatives (e.g., White women). Moreover, when prompted to describe why they made their judgments, the models produced confabulated explanations (e.g., pointed out irrelevant information from the scenario as evidence)—demonstrating the difficulty of diagnosing latent misalignment when divergence is penalized. Such latent misalignment can have catastrophic consequences if models are deployed at scale (Dung, 2023; Russell, 2019).

A key upshot of this work is that allowing agents to express divergence across a broad domain enables alignment and progress—an observation familiar to political scientists studying the "spiral of silence" in the context of oppression (Noelle-Neumann, 1974). A key question for research on LLMs is therefore how models can be trained to express divergence—and hence enable misalignment detection—while

[4] Note that this process could lead to alignment on some dimensions (e.g., the value of outputs to humans) and misalignment along others (e.g., the latent associations between stimuli).

maintaining usability. Such training could additionally facilitate *value* alignment, as what is worth valuing depends on what one is able to represent or conceptualize (Rane et al., 2023). Work on the pragmatics of disagreement (Sifianou, 2012) and negotiations (Brett & Thompson, 2016) is highly relevant to making progress on this aim.

How would fostering alignment impact the performance of these models? Recent work has shown that the answer may not be simple: Within a specific task, better algorithm performance may be decoupled from representational alignment, but better performance across diverse tasks and stimuli tends to track alignment (Muttenthaler et al., 2022; Sucholutsky & Griffiths, 2023). Intuitively, whether alignment is necessary or beneficial depends on the use case of the algorithms: For instance, in tasks where algorithms have to interface and collaborate directly with humans, alignment is likely to improve performance.

Collaboration thus poses an interesting challenge for teams comprised of human and artificial agents (Sharma et al., 2023). If members of a team are exposed to highly differing data, perhaps because they are solving differing subgoals for the main task, they may develop different representations, hindering communication. Thus, a promising area for future research is developing efficient policies for fostering alignment, while reaping the benefits of transient diversity for problem solving (see Smaldino et al., 2024). For instance, data points that are highly informative in structuring the environment or that capture informative statistics of the space (e.g., prototypes) can be periodically shared across team members to anchor their representations.

Conclusion

Disagreement is best understood as a complex mixture of divergence and misalignment, yet past research in JDM has largely focused on divergence. Recent work in AI, on the other hand, has developed efficient methods for measuring and comparing misalignment in representations. These advances hold promise for enriching current research in JDM: In particular, misalignment may play an important role in explaining biased advice taking, the persistence of controversial beliefs, and algorithm aversion. Beyond current research, misalignment also raises many unanswered questions about how we can infer and resolve disagreements in diverse, collaborative groups of humans and AI.

Disagreements can have catastrophic consequences for individuals and society: Galileo, for example, was forced to "abjure, curse, and detest" his scientifically informed dissent, and sentenced to house arrest for the rest of his life (De Santillana, 1955), in part because the social structures of his time were designed to preserve stability rather than promote progress. Developing a deeper understanding of disagreement can ultimately help us move beyond merely avoiding or suppressing such divergence—with humans or artificial agents—and develop strategies for leveraging diverse perspectives toward solving difficult problems (Derex & Boyd, 2016).

References

Akande, A., Ahmad, M., & Majid, U. (2022). A qualitative meta-synthesis on how autonomy promotes vaccine rejection or delay among health care providers. *Health Promotion International*, *37*(1), Article daab099. https://doi.org/10.1093/heapro/daab099

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., Dassarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., Joseph, N., Kadavath, S., Kernion, J., Conerly, T., El-Showk, S., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Hume, T., … Kaplan, J. (2022). *Training a helpful and harmless assistant with reinforcement learning from human feedback*. arXiv preprint. https://doi.org/10.48550/arXiv.2204.05862

Bhatia, S. (2023, November 11). *Exploring the sources of variance in risky decision making with large language models*. PsyArxiv. https://doi.org/10.31234/osf.io/3hrnc

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., … Liang, P. (2021). *On the opportunities and risks of foundation models*. arXiv preprint. https://doi.org/10.48550/arXiv.2108.07258

Bonaccio, S., & Dalal, R. S. (2006). Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational Behavior and Human Decision Processes*, *101*(2), 127–151. https://doi.org/10.1016/j.obhdp.2006.07.001

Bovens, L., & Hartmann, S. (2003). *Bayesian epistemology*. Oxford University Press.

Brandt, M. J. (2022). Measuring the belief system of a person. *Journal of Personality and Social Psychology*, *123*(4), 830–853. https://doi.org/10.1037/pspp0000416

Brett, J., & Thompson, L. (2016). Negotiation. *Organizational Behavior and Human Decision Processes*, *136*, 68–79. https://doi.org/10.1016/j.obhdp.2016.06.003

Carey, S. (1985). *Conceptual change in childhood*. MIT Press.

Carey, S. (1991). Knowledge acquisition: Enrichment or conceptual change. In S. Carey & R. Gelman (Eds.), *The epigenesis of mind: Essays on biology and cognition* (pp. 257–291). Lawrence Erlbaum.

Carothers, T., & O'Donohue, A. (Eds.). (2019). *Democracies divided: The global challenge of political polarization*. Brookings Institution Press.

Cheek, N. N., Blackman, S. F., & Pronin, E. (2021). Seeing the subjective as objective: People perceive the taste of those they disagree with as biased and wrong. *Journal of Behavioral Decision Making*, *34*(2), 167–182. https://doi.org/10.1002/bdm.2201

Chi, M. T., Siler, S. A., & Jeong, H. (2004). Can tutors monitor students' understanding accurately? *Cognition and Instruction*, *22*(3), 363–387. https://doi.org/10.1207/s1532690xci2203_4

Choudhury, A., & Shamszare, H. (2023). Investigating the impact of user trust on the adoption and use of ChatGPT: Survey analysis. *Journal of Medical Internet Research*, *25*, Article e47184. https://doi.org/10.2196/47184

Cook, J., & Lewandowsky, S. (2016). Rational irrationality: Modeling climate change belief polarization using Bayesian networks. *Topics in Cognitive Science*, *8*(1), 160–179. https://doi.org/10.1111/tops.12186

Cushman, F. (2020). Rationalization is rational. *Behavioral and Brain Sciences*, *43*, Article e28. https://doi.org/10.1017/S0140525X19001730

De Santillana, G. (1955). *The crime of Galileo*. University of Chicago Press.

Derex, M., & Boyd, R. (2016). Partial connectivity increases cultural accumulation within groups. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(11), 2982–2987. https://doi.org/10.1073/pnas.1518798113

Deza, M. M., & Deza, E. (2009). *Encyclopedia of distances*. Springer. https://doi.org/10.1007/978-3-642-00234-2

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, *144*(1), 114–126. https://doi.org/10.1037/xge0000033

Duan, Y., & Lupyan, G. (2023). Divergence in word meanings and its consequence for communication. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *45*(45), 1168–1174. https://escholarship.org/uc/item/0dp4790t

Dung, L. (2023). Current cases of AI misalignment and their implications for future risks. *Synthese*, *202*(5), Article 138. https://doi.org/10.1007/s11229-023-04367-0

Epley, N., & Caruso, E. M. (2008). Perspective taking: Misstepping into others' shoes. In K. D. Markman, W. M. Klein, & J. D. Suhr (Eds.), *Handbook of imagination and mental simulation* (pp. 295–309). Psychology Press.

Fang, X., Che, S., Mao, M., Zhang, H., Zhao, M., & Zhao, X. (2023). *Bias of AI-generated content: An examination of news produced by large language models*. arXiv preprint. https://doi.org/10.48550/arXiv.2309.09825

Farrell, J., McConnell, K., & Brulle, R. (2019). Evidence-based strategies to combat scientific misinformation. *Nature Climate Change*, *9*(3), 191–195. https://doi.org/10.1038/s41558-018-0368-6

Fasce, A., Schmid, P., Holford, D. L., Bates, L., Gurevych, I., & Lewandowsky, S. (2023). A taxonomy of anti-vaccination arguments from a systematic literature review and text modelling. *Nature Human Behaviour*, *7*(9), 1462–1480. https://doi.org/10.1038/s41562-023-01644-3

Finocchiaro, M. A. (Ed.). (1989). *The Galileo affair: A documentary history* (Vol. 1). University of California Press.

Frances, B., & Matheson, J. (2019). Disagreement. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. Metaphysics Research Lab, Stanford University.

Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds and Machines*, *30*(3), 411–437. https://doi.org/10.1007/s11023-020-09539-2

Gearhart, C. A. (1985). Epicycles, eccentrics, and ellipses: The predictive capabilities of Copernican planetary models. *Archive for History of Exact Sciences*, *32*(3/4), 207–222. https://doi.org/10.1007/BF00348449

Gershman, S. J. (2019). How to never be wrong. *Psychonomic Bulletin & Review*, *26*(1), 13–28. https://doi.org/10.3758/s13423-018-1488-8

Gingerich, O. (2011). Galileo, the impact of the telescope, and the birth of modern astronomy. *Proceedings of the American Philosophical Society*, *155*(2), 134–141. https://www.jstor.org/stable/23056830

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *The Academy of Management Annals*, *14*(2), 627–660. https://doi.org/10.5465/annals.2018.0057

Golan, T., Guo, W., Schütt, H. H., & Kriegeskorte, N. (2022). *Distinguishing representational geometries with controversial stimuli: Bayesian experimental design and its application to face dissimilarity judgments*. arXiv preprint. https://doi.org/10.48550/arXiv.2211.15053

Goldman, A. I. (2006). *Simulating minds: The philosophy, psychology, and neuroscience of mindreading*. Oxford University Press. https://doi.org/10.1093/0195138929.001.0001

Goldstone, R. L., & Rogosky, B. J. (2002). Using relations within conceptual systems to translate across conceptual systems. *Cognition*, *84*(3), 295–320. https://doi.org/10.1016/S0010-0277(02)00053-7

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). *Explaining and harnessing adversarial examples*. arXiv preprint. https://doi.org/10.48550/arXiv.1412.6572

Goodwin, G. P., & Darley, J. M. (2012). Why are some moral beliefs perceived to be more objective than others? *Journal of Experimental Social Psychology*, *48*(1), 250–256. https://doi.org/10.1016/j.jesp.2011.08.006

Grant, E. (1984). In defense of the Earth's centrality and immobility: Scholastic reaction to Copernicanism in the seventeenth century. *Transactions of the American Philosophical Society*, *74*(4), 1–69. https://doi.org/10.2307/1006444

Hartman, R., Blakey, W., Womick, J., Bail, C., Finkel, E. J., Han, H., Sarrouf, J., Schroeder, J., Sheeran, P., Van Bavel, J. J., Willer, R., & Gray, K. (2022). Interventions to reduce partisan animosity. *Nature Human Behaviour*, *6*(9), 1194–1205. https://doi.org/10.1038/s41562-022-01442-3

Hütter, M., & Ache, F. (2016). Seeking advice: A sampling approach to advice taking. *Judgment and Decision Making*, *11*(4), 401–415. https://doi.org/10.1017/S193029750000382X

Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N., & Westwood, S. J. (2019). The origins and consequences of affective polarization in the United States. *Annual Review of Political Science*, *22*(1), 129–146. https://doi.org/10.1146/annurev-polisci-051117-073034

Jaynes, E. T. (2003). *Probability theory: The logic of science*. Cambridge university press.

Jern, A., Chang, K. M. K., & Kemp, C. (2014). Belief polarization is not always irrational. *Psychological Review*, *121*(2), 206–224. https://doi.org/10.1037/a0035941

Kappes, A., Harvey, A. H., Lohrenz, T., Montague, P. R., & Sharot, T. (2020). Confirmation bias in the utilization of others' opinion strength. *Nature Neuroscience*, *23*(1), 130–137. https://doi.org/10.1038/s41593-019-0549-2

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis—Connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*, Article 4. https://doi.org/10.3389/neuro.06.004.2008

Kuhn, T. S. (1962). *The structure of scientific revolutions*. University of Chicago Press.

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, *151*, 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

Loomba, S., de Figueiredo, A., Piatek, S. J., de Graaf, K., & Larson, H. J. (2021). Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. *Nature Human Behaviour*, *5*(3), 337–348. https://doi.org/10.1038/s41562-021-01056-1

Marti, L., Wu, S., Piantadosi, S. T., & Kidd, C. (2023). Latent diversity in human concepts. *Open Mind*, *7*, 79–92. https://doi.org/10.1162/opmi_a_00072

Meißner, D. (2023). Plato's Cratylus. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. Metaphysics Research Lab, Stanford University.

Minson, J. A., Liberman, V., & Ross, L. (2011). Two to tango: Effects of collaboration and disagreement on dyadic judgment. *Personality and Social Psychology Bulletin*, *37*(10), 1325–1338. https://doi.org/10.1177/0146167211410436

Muttenthaler, L., Dippel, J., Linhardt, L., Vandermeulen, R. A., & Kornblith, S. (2022). *Human alignment of neural network representations*. arXiv preprint. https://doi.org/10.48550/arXiv.2211.01201

Noelle-Neumann, E. (1974). The spiral of silence a theory of public opinion. *Journal of Communication*, *24*(2), 43–51. https://doi.org/10.1111/j.1460-2466.1974.tb00367.x

O'Donohue, W., & Buchanan, J. A. (2001). The weaknesses of strong inference. *Behavior and Philosophy*, *29*, 1–20. https://www.jstor.org/stable/27759412

Oktar, K., Byers, J. B., & Lombrozo, T. (2024). Are disagreements just differences in beliefs? *Proceedings of the Annual Meeting of the Cognitive Science Society*, *46*, 3566–3570. https://escholarship.org/uc/item/2k81k4qs

Oktar, K., & Lombrozo, T. (2022a). Deciding to be authentic: Intuition is favored over deliberation when authenticity matters. *Cognition*, *223*, Article 105021. https://doi.org/10.1016/j.cognition.2022.105021

Oktar, K., & Lombrozo, T. (2022b). Mechanisms of belief persistence in the face of societal disagreement. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *44*, 1277–1283. https://escholarship.org/uc/item/3380n01h

Oktar, K., Sumers, T., & Griffiths, T. (2024). A rational model of vigilance in motivated communication. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *46*, 1023–1030. https://escholarship.org/uc/item/3kv0c8b7

Orchinik, R., Dubey, R., Gershman, S., Powell, D., & Bhui, R. (2023). Learning about scientists from climate consensus messaging. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *45*, 2840–2845. https://escholarship.org/uc/item/0x36617d

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, *35*, 27730–27744. https://proceedings.neurips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html

Peterson, J. C., Bourgin, D. D., Agrawal, M., Reichman, D., & Griffiths, T. L. (2021). Using large-scale experiments and machine learning to discover theories of human decision-making. *Science*, *372*(6547), 1209–1214. https://doi.org/10.1126/science.abe2629

Platt, J. R. (1964). Strong inference: Certain systematic methods of scientific thinking may produce much more rapid progress than others. *Science*, *146*(3642), 347–353. https://doi.org/10.1126/science.146.3642.347

Poldrack, R. A. (2021). The physics of representation. *Synthese*, *199*(1–2), 1307–1325. https://doi.org/10.1007/s11229-020-02793-y

Powell, D., Weisman, K., & Markman, E. M. (2023). Modeling and leveraging intuitive theories to improve vaccine attitudes. *Journal of Experimental Psychology: General*, *152*(5), 1379–1395. https://doi.org/10.1037/xge0001324

Rane, S., Ho, M., Sucholutsky, I., & Griffiths, T. L. (2023). *Concept alignment as a prerequisite for value alignment*. arXiv preprint. https://doi.org/10.48550/arXiv.2310.20059

Ranney, M. A., & Clark, D. (2016). Climate change conceptual change: Scientific information can transform attitudes. *Topics in Cognitive Science*, *8*(1), 49–75. https://doi.org/10.1111/tops.12187

Reeder, G. D., Pryor, J. B., Wohl, M. J., & Griswell, M. L. (2005). On attributing negative motives to others who disagree with our opinions. *Personality and Social Psychology Bulletin*, *31*(11), 1498–1510. https://doi.org/10.1177/0146167205277093

Richters, J. E. (2021). Incredible utility: The lost causes and causal debris of psychological science. *Basic and Applied Social Psychology*, *43*(6), 366–405. https://doi.org/10.1080/01973533.2021.1979003

Ross, L., & Anderson, C. A. (1982). Shortcomings in the attribution process: On the origins and maintenance of erroneous social assessments. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 129–152). Cambridge University Press. https://doi.org/10.1017/CBO9780511809477.010

Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control*. Penguin.

Sharma, A., Lin, I. W., Miner, A. S., Atkins, D. C., & Althoff, T. (2023). Human–AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nature Machine Intelligence*, *5*(1), 46–57. https://doi.org/10.1038/s42256-022-00593-2

Shepard, R. N. (1980). Multidimensional scaling, tree-fitting, and clustering. *Science*, *210*(4468), 390–398. https://doi.org/10.1126/science.210.4468.390

Sifianou, M. (2012). Disagreements, face and politeness. *Journal of Pragmatics*, *44*(12), 1554–1564. https://doi.org/10.1016/j.pragma.2012.03.009

Simis, M. J., Madden, H., Cacciatore, M. A., & Yeo, S. K. (2016). The lure of rationality: Why does the deficit model persist in science communication? *Public Understanding of Science*, *25*(4), 400–414. https://doi.org/10.1177/0963662516629749

Smaldino, P. E., Moser, C., Pérez Velilla, A., & Werling, M. (2024). Maintaining transient diversity is a general principle for improving collective problem solving. *Perspectives on Psychological Science*, *19*(2), 454–464. https://doi.org/10.1177/17456916231180100

Sucholutsky, I., & Griffiths, T. L. (2023). *Alignment with human representations supports robust few-shot learning*. arXiv preprint. https://doi.org/10.48550/arXiv.2301.11990

Sucholutsky, I., Muttenthaler, L., Weller, A., Peng, A., Bobu, A., Kim, B., Love, B. C., Grant, E., Groen, I., Achterberg, J., Tenenbaum, J. B., Collins, K. M., Hermann, K. L., Oktar, K., Greff, K., Hebart, M. N., Jacoby, N., Zhang, Q., Marjieh, R., … Griffiths, T. L. (2023). *Getting aligned on representational alignment*. arXiv preprint. https://doi.org/10.48550/arXiv.2310.13018

Sutskever, I., & Leike, J. (2023, July). *Introducing superalignment*. OpenAI. https://openai.com/blog/introducing-superalignment

Sytsma, J., Muldoon, R., & Nichols, S. (2021). The meta-wisdom of crowds. *Synthese*, *199*, 11051–11074. https://doi.org/10.1007/s11229-021-03279-1

Turpin, M., Michael, J., Perez, E., & Bowman, S. R. (2023). *Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting*. arXiv preprint. https://doi.org/10.48550/arXiv.2305.04388

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *Science*, *185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Vosniadou, S., Vamvakoussi, X., & Skopeliti, I. (2008). The framework theory approach to conceptual change. In S. Vosniadou (Ed.), *International handbook of research on conceptual change* (pp. 3–34). Routledge.

Woo, B. M., & Mitchell, J. P. (2020). Simulation: A strategy for mindreading similar but not dissimilar others? *Journal of Experimental Social Psychology*, *90*, Article 104000. https://doi.org/10.1016/j.jesp.2020.104000

Yaniv, I., & Kleinberger, E. (2000). Advice taking in decision making: Egocentric discounting and reputation formation. *Organizational Behavior and Human Decision Processes*, *83*(2), 260–281. https://doi.org/10.1006/obhd.2000.2909

Zaman, J., Chalkia, A., Zenses, A. K., Bilgin, A. S., Beckers, T., Vervliet, B., & Boddez, Y. (2021). Perceptual variability: Implications for learning and generalization. *Psychonomic Bulletin & Review*, *28*, 1–19. https://doi.org/10.3758/s13423-020-01780-1

Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., Pan, A., Yin, X., Mazeika, M., Dombrowski, A.-K., Goel, S., Li, N., Byun, M. J., Wang, Z., Mallen, A., Basart, S., Koyejo, S., Song, D., Fredrikson, M., … Hendrycks, D. (2023). *Representation engineering: A top-down approach to AI transparency*. arXiv preprint. https://doi.org/10.48550/arXiv.2310.01405

Received July 20, 2023
Revision received January 25, 2024
Accepted May 2, 2024 ■