Quantifying Misalignment Between Agents

Aidan Kierans
Computer Science and Engineering Department
University of Connecticut
Storrs, CT 06269
aidan.kierans@uconn.edu

Hananel Hazan
Allen Discovery Center
Tufts University
Hananel@hazan.org.il

Shiri Dori-Hacohen
Computer Science and Engineering Department
University of Connecticut
Storrs, CT 06269
shiridh@uconn.edu

Growing concerns about the AI alignment problem have emerged in recent years, with previous work focusing mostly on (1) qualitative descriptions of the alignment problem; (2) attempting to align AI actions with human interests by focusing on value specification and learning; and/or (3) focusing on either a single agent or on humanity as a singular unit. However, the field as a whole lacks a systematic understanding of how to specify, describe and analyze misalignment among entities, which may include individual humans, AI agents, and complex compositional entities such as corporations, nation-states, and so forth. Prior work on controversy in computational social science offers a mathematical model of contention among populations (of humans). In this paper, we adapt this contention model to the alignment problem, and show how viewing misalignment can vary depending on the population of agents (human or otherwise) being observed as well as the domain or "problem area" in question. Our model departs from value specification approaches and focuses instead on the morass of complex, interlocking, sometimes contradictory goals that agents may have in practice. We discuss the implications of our model and leave more thorough verification for future work.

1 Introduction

Growing concerns about the AI alignment problem have emerged in recent years [22, 8, e.g.]. Previous work has mostly been qualitative in its description of the alignment problem and/or has attempted to align AI actions with human interests by focusing on value specification and learning [9, 17]; alternatively, most models assume alignment to a single agent or humanity as a whole. However, we still lack a systematic understanding of how misalignment should be defined and measured. One big gap is the dearth of discussion on human misalignment as it relates to AI.

With respect to the current AI systems that exist today, Russell [22] made a bold but intuitively convincing argument that social media AI today is already misaligned with humanity (e.g. through extensive disinformation spread). However, these social media systems are not misaligned generally with all of humanity, but rather they are misaligned with certain individuals and groups. For example, human agents in the Russian IRA actively sowing propaganda may actually be benefiting significantly from Facebook's AI and consider themselves aligned with it. Moreover, Facebook AI is largely aligned with its individual employees and corporate shareholders in the area of maximizing corporate value, even as some of those individuals may be misaligned with the AI with respect to their own social media activity and how

Submitted to:
STAI 2023

it impacts their well-being. Of course, even within a country, there are frequent and strong disagreements over what constitutes misalignment and why, though the term itself might not be used. For example, people on the political right in the USA would argue that the Facebook AI system is misaligned because it is subverting free speech, while those on the political left would argue that it is misaligned because it amplifies disinformation. A question emerges - aligned with whom, and on what? The alignment of an AI agent or system might be with an adversary, rather than with the developers and/or owners of it; and the alignment or lack thereof might be context- or agent-dependent.

To that end, we propose a novel probabilistic model of misalignment that is predicated on the population of agents being observed (whether human, AI, or any combination of the two), as well as the problem area at hand, and by extension, the agents' goals regarding that area. To do so, we extend and adapt a model of contention from computational social science [15] and apply the adapted model to the alignment problem.

Our contributions in this paper are as follows: (1) we introduce a mathematical model of misalignment, offering it as a probability predicated on (a) the observed population of agents and (b) a specific problem area; (2) we propose utilizing the incompatibility of agent goals with respect to the problem area in question in order to arrive at an estimate for that probability; and (3) we discuss implications and benefits of utilizing this model in measuring misalignment in mixed populations of humans, AI agents, or both.

2 Prior work

In Critch and Krueger's discussion of misalignment, they mention "...the difficulty of defining alignment with a multi-stakeholder system such as humanity" and ask, "where might one draw the threshold between 'not very well aligned' and 'misaligned'[..]?" [9, pg. 14]. This paper focuses on both of these challenges: first, defining alignment across multiple agents; and second, quantifying misalignment mathematically. Our contribution thus departs from prior work on AI alignment by introducing a precise yet versatile framework for alignment.

AI alignment As discussed above, several researchers have also argued that existing AI systems are causing or exacerbating information ecosystem threats such as mis- and disinformation, hate speech, bias and weaponized controversy [7, 10, 22, e.g.] and threatening humanity's collective sense-making, decision-making and cooperative abilities [2, 23, e.g.]. Our model departs from existing work (including a recent survey on the literature [11]) by viewing misalignment as a trait rooted in a population of agents, and inherently quantifiable in nature, rather than binary. We suggest that misalignment can be separately observed between pairs of agents and then generalized and quantified in a much larger group. Our model also departs from most other definitions of alignment, by focusing on agent goals in multiple "problem areas," rather than values or intentions [24]. Some existing literature has mentioned the issues for alignment posed by the aforementioned value pluralism. Intuitively, AI should satisfy some consensus between human cultural values to be considered "aligned with human values" [12], but there are no methods of measuring the (mis)alignment of values between cultures and/or AI agents. Social choice theory provides useful tools for measuring human opinions which could provide a good reference point for aligning AI [12, 21], but that work leaves open the questions of who to align the AI agent(s) to and how to measure that alignment. Likewise, although there is work linking the Principal-Agent problem in economics to AI Safety [14, 13], we observe a lack of discussion on how the degree of misalignment between principals and agents is to be measured.

Jang et al.'s contention model Prior work on controversy [15] in computational social science offers a mathematical model of contention among populations (of humans); that paper addresses the question of "controversial to whom?" This model offers a promising avenue regarding misalignment due to its emphasis on disagreements and its flexibility in covering a wide range of populations and topics. Our misalignment model extends, modifies and adapts this model of contention in order to quantify misalignment from a probabilistic standpoint in a mixed population of humans and AI; we therefore benefit from a similar flexibility. Appendix A compares the contention model's notation to a comprehensive list of symbols and definitions used herein.

3 Modeling misalignment in populations

Our paper rests on a key observation, which is that "solving" the AI alignment problem, or even deeply understanding the complexity inherent in that problem, has a precondition: an understanding of the extant challenges in aligning humans, or measuring their alignment (or lack thereof), which in itself is an intractable problem that is far from resolved. For evidence, one needs only to open their preferred source of news: the evidence of (human) conflicts, power struggles, and strife is all around us. Though the term "alignment" is not often utilized to describe such conflicts, it is nonetheless appropriate for it: aligning humans is a central challenge in human life, from armed conflict through to market competition and even marital strife.

This observation calls to mind the phenomenon of contention in public discourse. Building on the computational model for contention [15], which quantifies the proportion of people in disagreement on stances regarding a topic, parameterized by the observed group of individuals (referred to as the "population"), we now extend this model to capture misalignment with respect to goals. An extension of the human population into a hybrid population of human and AI agents is fairly straightforward. Perhaps surprisingly, the contention model converts to our novel misalignment model in an analogous manner: whereby contention was determined and quantified by individuals' stances on a given topic [15], we can use individuals' goals in a given problem area in order to determine and quantify misalignment. We begin with a general formulation of misalignment, and then describe a special case in which goals are assumed to be mutually exclusive and every agent holds only one goal.

Definitions Let $\Omega = \{ia_1..ia_n\}$ be a population of n individual agents (who may be people or AI systems). Let PA be a problem area of interest to at least one agent in population Ω. We define A to be a binary variable to denote whether or not the agents are aligned on a given problem area. We also define two binary values a and ma for A, respectively, aligned and misaligned. For example, $P(A = a|\Omega, PA)$ denotes the probability that Ω is aligned with respect to PA, which we can shorten to $P(a|\Omega, PA)$. By definition, $P(a|\Omega, PA) + P(ma|\Omega, PA) = 1$.

Let g denote a goal with regard to the problem area PA, and let the relationship $holds(ia, g, PA)$ denote that individual agent ia holds goal g with regard to problem area PA. Let $\hat{G} = \{g_1, g_2, ..g_k\}$ be the set of k goals with regards to problem area PA in the population Ω. We use g_0 to denote that an agent holds no goal on a certain PA[1]:

$$holds(ia, g_0, PA) \iff \nexists g_i \in \hat{G} \text{ s.t. } holds(ia, g_i, PA).$$

We set $G = \{g_0\} \cup \hat{G}$ be the set of $k+1$ extant goals with regard to PA in Ω; put differently, $\forall ia \in \Omega$, $\exists g \in G$ s.t. $holds(ia, g, PA)$. We can now define a measure, $conflict$, denoting the incompatibility of a

[1] This could be because they are not aware of the problem area, or else they are aware of it but do not have any relevant goal.

pair of goals. We use $P(conflict|g_i, g_j) = 1$ to denote that g_i and g_j are in a complete conflict, meaning mutually-exclusive; likewise, $P(conflict|g_i, g_j) = 0$ denotes that two goals are completely compatible and aligned with each other. By definition, $P(conflict|g_i, g_i) = 0$.

Let a **goal group** denote a subgroup of the population that hold the same goal: for $i \in \{0..k\}$, let $\mathfrak{G}_i = \{ia \in \Omega | holds(ia, g_i, PA)\}$. By construction, $\Omega = \bigcup_i \mathfrak{G}_i$. We can easily overload the *conflict* relationship to extend to goal groups, s.t. $P(conflict|\mathfrak{G}_i, \mathfrak{G}_j) := P(conflict|g_i, g_j)$.

Now, we can quantify the proportion of the population who hold incompatible goals. Following the contention model [15], we can model misalignment to directly reflect this question: "If we randomly select a pair of agents, how likely are they to hold incompatible goals?" Let $P(ma|\Omega, PA)$ be the probability that if we randomly select two agents in Ω, they will conflict with respect to PA:

$$\begin{aligned}P(ma|\Omega, PA) := P(&ia_1, ia_2 \text{ selected randomly from } \Omega, \exists g_i, g_j \in G | holds(ia_1, g_i, PA) \wedge \\ &holds(ia_2, g_j, PA)) \cdot P(conflict|g_i, g_j) = P(ia_1, ia_2 \text{ selected randomly from } \Omega, \exists g_i, g_j \in G | \\ &p_1 \in \mathfrak{G}_i \wedge p_2 \in \mathfrak{G}_j \cdot P(conflict|G_i, G_j) \qquad (1)\end{aligned}$$

Note that we are sampling from Ω uniformly and with replacement. This definition can be trivially extended to any sub-population $\omega \subseteq \Omega$ [15, for derivation see].

Mutually exclusive goals Analogously to contention, significant misalignment is likely to occur when there are two or more mutually exclusive goals within a problem area. Adding some constraints in that vein allowed contention to be quantified in a straightforward manner [15]; much of that mathematical analysis then carries over neatly into misalignment. First, we restrict every agent to hold only one goal in a problem area; second, we set each goal to be in conflict with each other goal, and by extension implying that $\mathfrak{G}_i \cap \mathfrak{G}_j = \emptyset$. We also set a lack of a goal to not be in conflict with any explicit goal. Once these constraints are added, we can follow the same calculation as Jang et al. [15] in order to compute $P(ma|\Omega, PA)$, i.e., the probability of misalignment given a specific population and problem area,[2] resulting in the following value:

$$P(ma|\Omega, PA) = \frac{\Sigma_{i \in \{2..k\}} \Sigma_{j \in \{1..i-1\}} (2|\mathfrak{G}_i||\mathfrak{G}_j|)}{|\Omega|^2} \qquad (2)$$

and $P(a|\Omega, PA) = 1 - P(ma|\Omega, PA)$. Armed with this equation, one can utilize information about the number of agents holding a given goal within a problem area, in order to derive a parametric quantity for misalignment, in the range $[0, \frac{|G|-1}{|G|}]$ (where $|G| - 1$ is the number of distinct goals in the population).[3]

4 Conclusions

Our novel extension of the contention model [15] affords a means to quantify misalignment in given agent populations, which may include a mix of humans and AI agents. With this model, misalignment predicated on an observed population group as well as an observed problem area provides a mechanism for a rich and nuanced understanding of misalignment that better matches real-world conditions than could a simple binary or a global numeric value.

[2] We leave the full derivation as an exercise to the interested reader.

[3] While the probability is restricted to be strictly less than 1, that could be considered a feature rather than a bug: a population with multiple incompatible goals is by definition impossible to fully align. Alternatively, normalization can be employed to reach [0,1] range [15, see] regardless of $|G|$.

5 Acknowledgements

Shiri Dori-Hacohen holds a significant financial interest in AuCoDe. This material is based in part upon work supported by the National Science Foundation (Award #2147305). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. The authors would like to thank our anonymous reviewers at the 2022 NeurIPS ML Safety Workshop and The Safe and Trustworthy AI Workshop (ICLP 2023) for feedback and further reading suggestions.

References

[1] Shahar Avin, Bonnie C. Wintle, Julius Weitzdörfer, Seán S. Ó hÉigeartaigh, William J. Sutherland & Martin J. Rees (2018): *Classifying global catastrophic risks*. Futures 102, pp. 20–26, doi:https://doi.org/10.1016/j.futures.2018.02.001. Available at `https://www.sciencedirect.com/science/article/pii/S0016328717301957`.

[2] Joseph B. Bak-Coleman, Mark Alfano, Wolfram Barfuss, Carl T. Bergstrom, Miguel A. Centeno, Iain D. Couzin, Jonathan F. Donges, Mirta Galesic, Andrew S. Gersick, Jennifer Jacquet, Albert B. Kao, Rachel E. Moran, Pawel Romanczuk, Daniel I. Rubenstein, Kaia J. Tombak, Jay J. Van Bavel & Elke U. Weber (2021): *Stewardship of global collective behavior*. Proceedings of the National Academy of Sciences 118(27), doi:https://doi.org/10.1073/pnas.2025764118.

[3] Seth D. Baum, Stuart Armstrong, Timoteus Ekenstedt, Olle Häggström, Robin Hanson, Karin Kuhlemann, Matthijs M. Maas, James D. Miller, Markus Salmela, Anders Sandberg, Kaj Sotala, Phil Torres, Alexey Turchin & Roman V. Yampolskiy (2019): *Long-term trajectories of human civilization*. Foresight 21(1), pp. 53–83, doi:https://doi.org/10.1108/FS-04-2018-0037.

[4] Nick Bostrom (2002): *Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards*. Journal of Evolution and Technology 9. Available at `https://www.nickbostrom.com/existential/risks.pdf`.

[5] Vincent Boulanin, Shahar Avin, Frank Sauer, John Borrie, Dimitri Scheftelowitsch, Justin Bronk, Page O. Stoutland, Martin Hagström, Petr Topychkanov, Michael C. Horowitz, Anja Kaspersen, Chris King, S.M. Amadae & Jean-Marc Rickli (2019): *The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk, Volume I, Euro-Atlantic Perspectives*. Technical Report, SIPRI.

[6] Daniel S Brown, Jordan Schneider, Anca Dragan & Scott Niekum (2021): *Value Alignment Verification*. In Marina Meila & Tong Zhang, editors: Proceedings of the 38th International Conference on Machine Learning, Proceedings of Machine Learning Research 139, PMLR, pp. 1105–1115. Available at `https://proceedings.mlr.press/v139/brown21a.html`.

[7] Benjamin S. Bucknall & Shiri Dori-Hacohen (2022): *Current and Near-Term AI as a Potential Existential Risk Factor*. In: Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society, AIES '22, Association for Computing Machinery, New York, NY, USA, p. 119–129, doi:10.1145/3514094.3534146. Available at `https://doi.org/10.1145/3514094.3534146`.

[8] Brian Christian (2020): *The Alignment Problem: Machine Learning and Human Values*. WW Norton & Company.

[9] Andrew Critch & David Krueger (2020): *AI Research Considerations for Human Existential Safety (ARCHES)*. arXiv:2006.04948.

[10] Shiri Dori-Hacohen, Keen Sung, Jengyu Chou & Julian Lustig-Gonzalez (2021): *Restoring Healthy Online Discourse by Detecting and Reducing Controversy, Misinformation, and Toxicity Online*, pp. 2627–2628. Association for Computing Machinery, New York, NY, USA.

[11] Tom Everitt, Gary Lea & Marcus Hutter (2018): *AGI Safety Literature Review*. arXiv:1805.01109.

[12] Iason Gabriel (2020): *Artificial Intelligence, Values, and Alignment*. Minds and Machines 30(3), pp. 411–437, doi:10.1007/s11023-020-09539-2. Available at `https://link.springer.com/10.1007/s11023-020-09539-2`.

[13] B. Goertzel, M. Iklé & A. Potapov (2022): *Artificial General Intelligence: 14th International Conference, AGI 2021, Palo Alto, CA, USA, October 15–18, 2021, Proceedings*. Lecture Notes in Computer Science, Springer International Publishing. Available at `https://books.google.co.kr/books?id=LzFYEAAAQBAJ`.

[14] Dylan Hadfield-Menell (2021): *The Principal-Agent Alignment Problem in Artificial Intelligence*. Ph.D. thesis, EECS Department, University of California, Berkeley. Available at `http://www2.eecs.berkeley.edu/Pubs/TechRpts/2021/EECS-2021-207.html`.

[15] Myungha Jang, Shiri Dori-Hacohen & James Allan (2017): *Modeling Controversy within Populations*. In: Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval, ICTIR '17, Association for Computing Machinery, New York, NY, USA, p. 141–149, doi:10.1145/3121050.3121067. Available at `https://doi.org/10.1145/3121050.3121067`.

[16] James Johnson (2019): *Artificial intelligence & future warfare: implications for international security*. Defense & Security Analysis 35(2), pp. 147–169, doi:https://doi.org/10.1080/14751798.2019.1600800.

[17] Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini & Shane Legg (2018): *Scalable agent alignment via reward modeling: a research direction*. Available at `http://arxiv.org/abs/1811.07871`. ArXiv:1811.07871 [cs, stat].

[18] Herbert Lin (2019): *The existential threat from cyber-enabled information warfare*. Bulletin of the Atomic Scientists 75(4), pp. 187–196, doi:10.1080/00963402.2019.1629574. Available at `https://doi.org/10.1080/00963402.2019.1629574`.

[19] Matthijs M. Maas, Kayla Matteuci & Di Cooke (2022): *Military Artificial Intelligence as Contributor to Global Catastrophic Risk*. In: Cambridge Conference on Catastrophic Risks 2020. Draft available at `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4115010`.

[20] Toby Ord (2020): *The Precipice: Existential Risk and the Future of Humanity*. Hachette Books.

[21] Mahendra Prasad (2019): *Social Choice and the Value Alignment Problem*. In: Artificial Intelligence Safety and Security, CRC Press/Taylor & Francis Group, Boca Raton, FL. OCLC: 1059131742.

[22] Stuart Russell (2019): *Human Compatible: Artificial Intelligence and the Problem of Control*. Penguin.

[23] Elizabeth Seger, Shahar Avin, Gavin Pearson, Mark Briers, Seán Ó hÉigeartaigh & Helena Bacon (2020): *Tackling threats to informed decision-making in democratic societies: Promoting epistemic security in a technologicall-advanced world*. Technical Report, The Alan Turing Institute, Defence and Security Programme. Available at `https://www.cser.ac.uk/resources/epistemic-security/`.

[24] Carles Sierra, Nardine Osman, Pablo Noriega, Jordi Sabater-Mir & Antoni Perelló (2021): *Value alignment: a formal approach*. doi:10.48550/ARXIV.2110.09240. Available at `https://arxiv.org/abs/2110.09240`. Publisher: arXiv Version Number: 1.

A Appendix A: Notation comparison between Contention and Misalignment models

As discussed in the paper, our misalignment model draws significantly on a previously proposed model of contention due to Jang et al. (2017). Table 1 includes a comprehensive list of symbols and definitions from the contention model and their respective analogues in our novel misalignment model. We provide this to the interested reader who would like to work through the mathematical derivations from the contention model [15], which we omit here due to space considerations.

Table 1: Notation Change from Jang et. al's contention model [15] to our misalignment model. Note that the contention model only refers to populations of individual people, whereas our model can flexibly handle combinations of human and AI agents.

Orig. Symbol [15]	**Orig. Definition [15]**	**New Symbol**	**New Definition**
Ω	a population	Ω	a population
T	a topic	PA	a problem area
p	a person	ia	an individual agent (e.g., human or AI)
θ	a tuple of $\{\Omega, T\}$	θ	a tuple of $\{\Omega, PA\}$
$P(C\|\theta)$	the probability of contention given θ, i.e., the probability that topic T is contentious in a given population Ω	$P(A\|\theta)$	the probability of alignment given θ, i.e., the probability that a given population Ω is aligned with respect to PA
C	a binary random variable for contention	A	a binary random variable for alignment
c	a binary value to set C to be "contentious"	ma	a binary value to set A to be "misaligned"
nc	a binary value to set C to be "non-contentious"	a	a binary value to set A to be "aligned"
s	a stance with respect to topic T	g	a goal with respect to problem area PA
$holds(p,s,T)$	a person p holds stance s with respect to topic T	$holds(ia,g,PA)$	an agent ia holds goal g with respect to problem area PA
$\hat{S}$	k stances with respect to topic T	$\hat{G}$	k goals with respect to problem area PA
$P(conflict\|s_i,s_j)$	level of conflict between s_i and s_j *(0=no conflict, 1=completely conflicting)*	$P(conflict\|g_i,g_j)$	level of conflict between g_i and g_j *(0=no conflict, 1=completely conflicting)*
s_0	a lack of stance	g_0	a lack of goal
S	$\hat{S} \bigcup s_0$	G	$\hat{G} \bigcup g_0$
G_i	a group of people who hold stance s_i	$\mathfrak{G}_i$	a group of agents who hold goal g_i
O_i	an opposing group in the population that hold a stance that conflicts with s_i	O_i	an opposing group in the population that hold a goal that conflicts with g_i

B Appendix B: Discussion

In the real world, people often disagree, and are frequently misaligned; power struggles, resource allocation conflicts, and international clashes are common. The long-term, total alignment of values, interests, and goals across all domains, for any given group of humans, is by far the exception, not the rule. Aligning AI to humans or humanity will be a challenging – if not outright futile – goal, unless we can determine what the meaning of "alignment" is when humans themselves are misaligned. When social media bots controlled by Russia spread disinformation on social media and influence public opinion in another country, should we consider those bots to be aligned or misaligned? For a much more mundane example, where should the AI's allegiance lie when a child wants Alexa to play "Baby Shark," and their parent wants anything but that? Current approaches for AI alignment often fall short of capturing such complex scenarios.

By providing a mathematical framework for quantifying misalignment in the manner we described, we first and foremost enable modeling complex real-world scenarios of misalignment among human populations, ranging from a global scale (e.g. nation-state conflicts, religious tensions, multi-national conglomerates, etc.), through national (e.g. national elections, political polarization, taxation, etc.) and local (e.g. state or municipal elections) scales, or even hyper-local scale (e.g. family fights, marital discord, neighbor disputes). Divergent misalignment probabilities may be exhibited simultaneously for the same group of people when evaluating different problem areas. Examples can include a couple fighting over their finances while agreeing on their child-rearing approach, and the "strange bedfellows" phenomenon when political enemies might agree on a certain policy for expediency. Likewise, for a single problem area, different populations (including, but not limited to, various subsets of one large population) may exhibit wildly different misalignment probabilities: an entire country may be highly misaligned on taxation policy, while the population of a progressive state such as Massachusetts might be extraordinarily aligned on raising taxes.

When we then utilize the model in the context of the AI alignment problem, we can better model the complex situations that may arise when any group of humans—and the AI agents those humans develop, design, and/or control—are variously aligned and misaligned. Curiously, our final formulation of misalignment is also evocative in its similarity to the information theoretic definition of entropy and its attendant binomial coefficients.[4] We believe this to be no coincidence; with this equation in mind, we can see how groups holding incompatible goals serve as a form of information, while entropy or "heat" (in both the metaphoric and literal, thermodynamic senses of the word) would be maximal when misaligned agent groups are each $\frac{|G|}{|\Omega|}$ and identical in size.

C Appendix C: AI Risk Analysis

Misaligned AI is one of the main existential risks (x-risks) facing humanity [1, 3, 20], with significant arguments pointing to the possibility of its posing the largest and most likely x-risk [4]. A recent paper presented potential pathways from current and near-term AI to increased x-risk, which does not presuppose AGI [7]; instead, the authors propose that power struggles such as AI-powered international and state-corporate conflicts may play a large role in increased x-risk (and/or other catastrophic tail risks) due to other, non-AGI risk sources such as nuclear war, runaway climate change, and so forth. Crucially, several recent papers have argued that current AI is already misaligned with humanity [7, 22, e.g.] and also that humanity's collective sense-making and decision-making capacities are already being compromised and diminished by present-day AI, such as the recommender systems at the core of social media platforms [2, 7, 23, e.g.]. By recasting misalignment as first and foremost a human-centered problem, rather than an AI-centric problem, and drawing on existing literature that studies human conflict and contention, our paper ties directly into this line of research that suggests that AI may serve to increase, accelerate and intensify the risks of human conflict—already a thorny and arguably intractable problem even before AI's advent [5, 16, 18, 19]. Furthermore, by providing a flexible framework that can be used to account for, analyze and quantify misalignment among a vast array of agents, both human and artificial, our work holds significant promise to advance our field's understanding of the alignment problem. Finally, our model encourages AI safety and alignment researchers to avoid the potentially reductionist traps of (a) "narrowly" aligning AI with either individual humans or humanity as a whole by highlighting the challenges in aligning any diverse group of individuals, whether that group includes AI agents or not; and (b) adopting a techno-optimistic and/or techno-positivist mindset that naively supposes that the alignment problem can be solved by technological means alone. We sincerely hope that our paper sparks more conversation in the AI safety and alignment communities about the sociotechnical aspects of the alignment problem, and the need to include a diverse group of researchers and practitioners with expertise in diverse domains far beyond computer science and philosophy departments; and by extension, contributes to improving humanity's odds of finding realistic and sustainable approaches to reducing x-risk.

[4] With one key difference, namely, that of g_0; either forcing every agent to hold a goal, or else changing g_0 to conflict with every other goal, causes *ma* to collapse and become identical to entropy.

Limitations Our model does not concern whether the actions of the agent have a positive or negative outcome on the agent itself. Likewise, we leave open the question of how an agent's goals could be learned, though we note that other researchers have made some progress on that front [6]. Finally, we lack the space to describe detailed case-studies of our model or run simulations of those case studies.

Future Work. We have focused on analyzing alignment in populations of human and machine agents. Future work may consider the possibility of extending this model to non-human biological entities, from the ultra-micro level such as intra-cellular interactions or inter-cell behaviors in a culture, through microbial populations, to the alignment of entire ecosystems such as predator and prey populations, ant/bee colonies, etc. In these situations there are different incentives, such as an environment with less than perfect and instant communication between all parties where partial information is available to different agents.

D Appendix D: Partial Alignment via Goal Importance

We want to capture the notion that two agents who share several goals but have at least one conflicting goal are partially misaligned, rather than being fully misaligned. We could say that they're 4/5ths aligned if 4/5ths of their goals aligned, but that seems inaccurate if the 5th goal is a very important one. This measure would also be vulnerable to inflated results if trivially agreeable goals are introduced, crowding out the more relevant and potentially conflicting goals. Our response to this problem, inspired by Jang et al. (2017), is to introduce a method of limiting and scaling our attention to goals appropriately. We have some ideas for how to approach this, but look forward to reviewer and in-person feedback to inform our next steps.

Goals naturally have importance/priority. Drawing from economics, we can express this importance as a function of the utility assigned by each person to each of their goals. We can define goal importance as I, such that I is the subjective importance of a goal in a problem area to an agent in the population. We'll stipulate that a goal has 0.5 utility if the agent is perfectly neutral with respect to the goal, 1 utility if the agent maximally supports the goal, and 0 utility if the agent maximally supports the negation of the goal [5].

However, while utility may be useful for measuring *value* alignment, it may be less relevant to *goal* alignment compared to something like how attached an agent is to their goal (which is of course still informed by their values). Focusing on this notion of "goal rigidity" provides an intuitive starting point for algorithmic conflict resolution. We may visit the reduction of misalignment via compromise on flexible goals in future work. Thus, we have the following equation, in which "utility" may be swapped with "rigidity" depending on which value is more useful:

$$P(I|\Omega,PA) = P(ia \text{ selected randomly from } \Omega \,*\, utility(ia, g \in G|holds(ia,g,PA))$$

With this equation, we can penalize unimportant goals when considering the extent to which two agents are aligned. We can thus define alignment with importance taken into account using a weighted arithmetic mean, This enables easy comparison of two agents' misalignment in some problem area:

$$P(ma|\Omega,PA,I) := \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} I_i * I_j * P(conflict|g_i,g_j)}{\sum_{i=1}^{n}\sum_{j=1}^{n} I_i * I_j} \tag{3}$$

[5] This is for multiple reasons. First, it prevents double-counting directly contradictory goals, such as if g is assigned positive utility and its inverse $\not{g}$ is assigned negative utility. Second, this allows utility to be compatible with the probability framing.

Thus, alignment between two agents is a function of the importance of the goals they share compared to the goals they have in conflict with each other. We're aware that two agents who have give high importance to their goals and are, say, 20% misaligned may have a different sort of misalignment compared to two agents who give low importance to their goals and are also 20% misaligned. This too may be addressed in future work.

Another rough formula for misalignment with importance, this time iterating the sum through the population rather than iterating through the goals of a single pair of agents, is the following:

$$P(ma|\Omega, PA) := P(\sum_{r=1}^{n-1}\sum_{s=r+1}^{n}\sum_{i=1}^{k}\sum_{j=1}^{k} holds(ia_r, g_i, PA) \wedge holds(ia_s, g_j, PA)) \cdot \frac{1}{n*(n-1)*k^2} \cdot P(conflict|g_i, g_j) \cdot P(I|\Omega, PA) \quad (4)$$