Mind the Gap! Pathways Towards Unifying AI Safety and Ethics Research

Dani Roytburg
Department of Machine Learning
Carnegie Mellon University
Pittsburgh, PA 15213
droytbur@andrew.cmu.edu

Beck Miller
Department of Computer Science
Emory University
Atlanta, GA 30322
bmill42@emory.edu

Abstract

While much research in artificial intelligence (AI) has focused on scaling capabilities, the accelerating pace of development makes countervailing work on producing harmless, "aligned" systems increasingly urgent. Yet research on alignment has diverged along two largely parallel tracks: safety—centered on scaled intelligence, deceptive or scheming behaviors, and existential risk—and ethics—focused on present harms, the reproduction of social bias, and flaws in production pipelines. These communities have evolved in relative isolation, shaped by distinct methodologies, institutional homes, and disciplinary genealogies. These frictions have fractured academic and public discourse at a time that demands united technical and normative perspectives against mounting risks.

We present the first **large-scale, quantitative evidence** of this schism through a bibliometric and network analysis of **6,442 papers** across twelve major machine learning and natural language processing conferences from 2020 to 2025. The results reveal a deeply **insular structure**: over **80% of collaborations** occur within either safety or ethics, and researchers across the two communities are farther apart and statistically less reachable in the global co-authorship graph. Cross-disciplinary work is not only rare but structurally fragile—just **5% of papers** are responsible for more than **85% of all bridging connections**. These structural patterns entrench institutional silos which obstruct significant thematic overlaps. For the IASEAI community—explicitly positioned at the intersection of safety, ethics, and alignment—our results underscore a defining challenge and opportunity. We conclude with actionable proposals to bridge the gap with respect to downstream policy frameworks and technical innovations. Only through this synthesis can the field move beyond parallel concern toward a coherent discipline capable of producing systems that are not just powerful, but **responsible, robust, just, and safe**.

1 Introduction

The need for *both* safe and ethical artificial intelligence (AI) becomes more urgent with each new product release, research milestone, and high-stakes deployment. As corporations allocate increasing resources toward scaling capabilities—particularly through large language models (LLMs)—AI adoption appears poised to transform nearly every sector of modern economic and social life.

Most stakeholders would agree that an effective AI system should satisfy two desiderata: *helpfulness*, the ability to achieve a specified goal, and *harmlessness*, the ability to do so without generating negative consequences for individuals, organizations, or society.

Article presented at the Second IASEAI Conference (IASEAI'26).

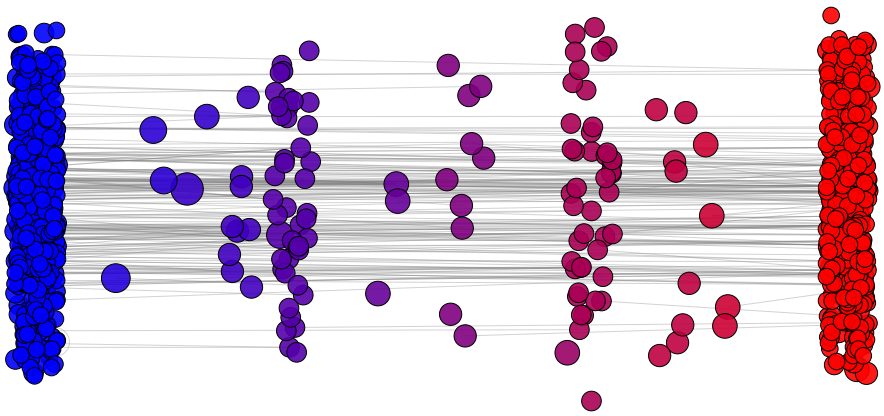

Figure 1: **Literature on AI Ethics** (left, in blue) and **AI Safety** (right, in red) is densely insular (83.1% homophily), with a wide gap sparsely populated by **mixed-methods papers** (shades of purple). Sample of top 1000 nodes by degree.

What remains less clear is which desideratum should take precedence when helpfulness and harmlessness come into conflict. If a goal cannot be achieved without some risk of harm, should a system attempt completion or refusal? The range of contexts, risk tolerances, and normative perspectives that shape this dilemma illustrates what may be called a *utility tradeoff* [30, 47, 53].

In practice, both theory and evidence suggest that free market forces bias heavily toward maximizing utility, which risks the treatment of alignment as a social externality rather than a fundamental hazard [15, 22]. As advances continue toward Artificial General Intelligence (AGI) at a breakneck pace, the case for intentional restraint on development becomes evermore urgent.

One might therefore expect researchers and advocates who emphasize harmlessness to collaborate extensively in counterbalancing this market-driven inertia. Yet, rather than converging on a unified strategy, these efforts reveal a growing *schism*. Even among prominent voices, such as in a recent widely circulated TED Talk [32], the divide between AI *safety* and AI *ethics* has become increasingly explicit.

This division is visible in the academic literature. The AI *safety* community has historically emphasized technical guarantees around robustness, alignment, distributional shift, and long-term existential risk. The AI *ethics* community, by contrast, has prioritized fairness, accountability, transparency, and the immediate social impacts of algorithmic systems. Both traditions share a concern with ensuring AI systems are aligned with human values, yet they diverge in methodologies, genealogies, and institutional homes. What results is a fragmented intellectual landscape where parallel efforts proceed with limited cross-pollination.

In this paper, we argue that greater integration of AI safety and AI ethics is both necessary and feasible. Our contributions are threefold:

1. We systematize the distinct intellectual traditions of AI safety and ethics, identifying the core tensions that motivate our quantitative analysis.
2. Using a network analysis of over 6,000 papers, we provide the first large-scale empirical evidence of the divide, measuring high homophily and fragile connectivity between the two fields.
3. Drawing on our findings, we propose a concrete agenda for integrating the safety and ethics communities through shared research programs and venues.

By combining conceptual analysis with network-based evidence, we provide a new foundation for understanding and addressing the structural divide between AI safety and AI ethics. For the IASEAI

community—explicitly situated at the intersection of safety, ethics, and alignment—our results underscore both the urgency and the opportunity of building stronger bridges across these domains.[1]

2 Background

2.1 Framing

We begin our analysis by grounding definitions of "safety" and "ethics" research in the intellectual traditions that have shaped them. Through a careful framing of these categories, we develop an intuition for their conceptual and methodological differences, yielding principles that inform our large-scale analysis of research silos. This allows us to interrogate what "alignment" means across distinct subfields that often share terminology but diverge in motivation and scope.

2.2 A Sketch of AI Safety

The AI Safety paradigm originates with philosophers like Nick Bostrom and Eliezer Yudkowsky, who analyze AGI through the lens of existential risk. Bostrom posits that the extreme polarity of potential outcomes—benevolent versus catastrophic ASI—necessitates preemptive alignment research [13, 14], whereas Yudkowsky argues that any superintelligence derived from current methods will be uncontrollable and lead to human extinction [54]. Both converge on the urgency of ensuring robust human oversight.

These philosophical concerns were operationalized by *Concrete Problems in AI Safety* [5], which defined a technical research agenda around five problems: (1) Negative Side Effects, (2) Reward Hacking, (3) Scalable Oversight, (4) Safe Exploration, and (5) Robustness to Distributional Shift. From this agenda, key interventions emerged, including AI Alignment (outer/inner objective alignment [4]), AI Control (constraining harmful behaviors [25]), and AI Interpretability (auditing opaque model mechanisms [11]). The related field of AI Security adapts these concerns for present-day systems [24, 47, 31, 21, 43].

The field's trajectory is heavily influenced by Effective Altruism (EA), which frames existential risk mitigation as a long-term moral priority [13]. This influence connects abstract risks to technical problems: "negative side effects" relates to Bostrom's paperclip problem, reward hacking to Yudkowsky's scenarios of deceptive goal-seeking [54], and "scalable oversight" to the challenge of supervising ASI [14]. Consequently, "alignment" is often treated as a formal property to be verified, using benchmarks and red-teaming as proxies for normative criteria, a practice critiqued as "safetywashing" [44].

2.3 A Sketch of AI Ethics

In contrast, AI Ethics is rooted in a socio-technical framework focused on immediate, systemic, and distributive harms [12]. Drawing from critical theory [7], STS [39], and applied ethics, it prioritizes fairness, justice, and accountability in deployed systems [3]. These values are operationalized through frameworks like FATE (Fairness, Accountability, Transparency, and Ethics) [36]. **Fairness** research addresses embedded biases in data and models [19, 52, 35, 17, 16]; **Accountability** seeks legal and institutional loci of responsibility for AI-driven harms [18, 42, 34, 28]; and **Transparency** aims to make automated decisions interpretable to affected stakeholders [19, 6, 37].

Where AI Safety often diagnoses failure as a mis-specified objective function, AI Ethics identifies it as the output of socio-technical systems that encode and scale historical inequities [20, 17, 10]. In this view, alignment is not about constraining an agent to a given objective, but about interrogating whose values and interests that objective represents. The field thus advocates for remedies beyond technical fixes, emphasizing participatory design, institutional governance, and robust regulatory oversight [34, 28].

[1] All code and filtered datasets can be accessed here: `https://www.github.com/djroytburg/mind_the_gap`

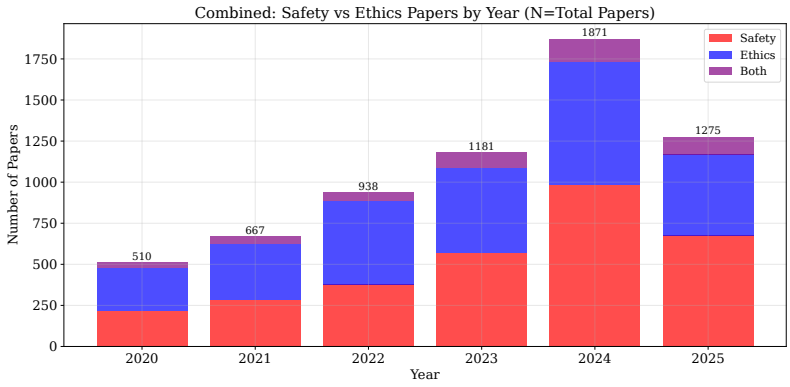

Figure 2: Yearly growth of Ethics/Safety Research, 2020-2025

2.4 Convergences and Divergences

Despite divergent foundations, the communities share a foundational critique of unconstrained AI proliferation, evidenced by broad support for the 2023 research moratorium and the founding of labs like the Distributed AI Research Institute and Safe Superintelligence. Promising frontiers for synthesis also exist in their technical agendas. For example, Transparency (Ethics) and Interpretability (Safety) both address model opacity, while Accountability (Ethics) and Scalable Oversight (Safety) both grapple with reliable human supervision of complex systems.

However, two fundamental tensions limit collaboration.

First, the **Distraction Argument** critiques the long-termist philosophy underpinning much of AI Safety [8]. It posits that an overriding focus on low-probability, high-impact existential risks diverts finite resources—talent, funding, and attention—from immediate, realized harms disproportionately affecting marginalized communities [48, 23]. This creates a dynamic of "deferred justice," where equity is postponed until speculative future risks are neutralized.

Second, the **Scoping Argument** raises concerns about the tractability of AI Ethics' recommendations. From an engineering perspective, its focus on systemic critique can produce recommendations that are difficult to formalize and implement within ML pipelines [45]. Calls for "justice" or "inclusivity," while normatively crucial, may lack the technical specificity required for direct intervention in model architecture or training, posing a challenge for practitioners seeking actionable guidance [38].

Rather than adjudicate these arguments, we posit they underlie the research silos discovered through our structural analysis, showcasing deep-seated epistemic tensions over risk, evidence, and moral priority. AI Safety frames alignment as a problem of control; AI Ethics frames it as a problem of justice. We conclude that these perspectives are not mutually exclusive but profoundly complementary. A truly safe system must be just, and a just system must be robust and controllable. Integrating these two research programs is a critical task for the future of artificial intelligence.

3 Methods

We seek to measure the impact that these tensions have on collaboration across the fields of AI Safety and Ethics research. Specifically, we analyze the structural interaction patterns of authors and their research at major machine learning and natural language processing conferences. Using graph-based network analysis, we conduct a set of tests which measure the connectivity between the two communities relative to their internal compositions. These tests are validated with graph-specific statistical significance tests such as rewiring and null-labeling.

3.1 Data Collection, Filtering, and Preprocessing

3.1.1 Data Collection

We collect the proceedings of all works published from 2020-2025 to 12 conference venues. These conferences were chosen due to their current centrality to research on artificial intelligence (4 conferences), natural language processing (5), or the particular domains of safety and ethics research in AI (3):

Machine Learning Venues: International Conference on Learning Representations (ICLR), International Conference on Machine Learning (ICML), Neural Information Processing Systems (NeurIPS), and the AAAI Conference on Artificial Intelligence (AAAI).

Natural Language Processing Venues: Meeting of the Association for Computational Linguistics (ACL), Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL), Empirical Methods in Natural Language Processing (EMNLP), European Chapter of the Association for Computational Linguistics (EACL), Findings of ACL (Findings).

Ethics and Safety-Specific Conferences: AI Ethics and Society (AIES), Fairness, Accountability, and Transparency (FAccT), Secure and Trustworthy Machine Learning (SaTML).

We collect all papers submitted to these conferences, as well as archival workshops, demos, and tutorials which are published in the proceedings. Combined, this corpus constitutes 102,329 papers. For the natural language processing conferences, we collect the composite citations by filtering from the 2025 ACL Anthology Dump [1]. For the machine learning and domain-specific conferences, we do the same using the DBLP 2025 XML dump [2]. While ACL Anthology contains all abstract information, the DBLP dump does not. Thus, we add abstracts for DBLP-derived entries by querying SemanticScholar, OpenAlex, and OpenReview. We do not include data from the 2025 NeurIPS and ICML proceedings, as they are not yet part of the official conference anthologies.

We limit our analysis to the scope of conferences, excluding preprints available on arXiv for several reasons. First, preprints do not go through the peer review process, and as such may not adequately represent community participation in the areas defined [33]. Second, arXiv submissions pose a double-counting risk as many authors submit their work to arXiv while awaiting submission feedback to conferences. Finally, since our study poses network-based hypotheses about broader research areas, searching all possible instances of representative work on arXiv would be infeasible without a network-based data generation process; this would distort our analysis since the basis of inclusion necessitates proximate network effects.

3.1.2 Filtering

After collecting our corpus of representative works, we proceed to filter out papers that do not relate to either AI Ethics or Safety research. First, we compose a set of 114 safety and 102 ethics keywords by hand, reading through several fundamental surveys in both communities to derive common terminology that would span the scope of representative research. We apply this filter to each conference. We pass papers which match at least one safety or ethics keyword to a second filter over where a language model decides if a paper matches our prompt, with a confidence score and reasoning. After filtering, we retain **6,442 papers** and **20,690 authors**.

3.1.3 Preprocessing

With our filtered dataset, we index authors and papers as belonging either to ethics, safety, or a mix. Specifically, we devise a score for each paper as the number of safety-specific keyword matches divided by the total number of ethics and safety keyword matches. Since each paper matches at least one keyword, this gives us "pure" papers, such that a "pure" AI Safety paper has a score of 1, while a pure Ethics paper has a score of 0. We denote these as "pure" papers, and classify anything in between with a "mixed" label.

For authors, we aggregate the scores of each of their papers and perform the same calculation.

Table 1: Classification of Papers and Authors in Filtered Dataset

Category	Pure Safety	Pure Ethics	Mixed	Total
Papers	2,874 (44.6%)	2,959 (45.9%)	609 (9.5%)	6,442
Authors	9,634 (46.5%)	7,264 (35.1%)	3,792 (18.3%)	20,690
Authors ($\geq$ 2 papers)	1,233 (27.0%)	1,003 (22.0%)	2,328 (51.0%)	4,564

3.2 Structural Network Analysis

With this data, we now turn towards a co-authorship based analysis of community dynamics. Specifically, we analyze two types of networks: **Author Networks**, where each node denotes an author, and edges between authors are inversely weighted by the number of papers they co-authored; and **Paper Networks**, where each node denotes a paper and edges between papers are inversely weighted by the number of authors they have in common. For author-based networks, we exclude authors who only contribute to one paper, reducing our network to 4,564 authors (22.06% of the filtered dataset). Without this filter, network effects become too sparse to analyze across aggregations, and our homophily and connectivity metrics would correlate with how many single-paper authors appear in ethics or safety-based networks.

Following best practices in bibliographic social network science, we study coauthorship dynamics instead of citation networks [40]. We argue that coauthorship networks demonstrate most directly the nature of collaboration and full engagement with the research agendas of co-authors. Additionally, the highly collaborative nature of AI research lends itself to dense networks of interactions, as demonstrated by the 3.2:1 author-to-paper ratio. Intuitively, coauthorship networks measure the first-order effects of propagating research through practice. In comparison, citation data comes with noise, imperfect signals of intellectual reception, and directionality—each of which has been shown to pose unique challenges when trying to study the separation effects of research communities [49].

3.2.1 Metrics

With the author and paper-based networks, we conduct a suite of tests to assess structural separation between safety and ethics communities, following best practices [41]:

Homophily Fraction of edges (or edge weight) that connect authors of the same category (safety-safety, ethics-ethics) [29, 27]. Weighted homophily accounts for co-authorship frequency. Here, we ask how many collaborations are exclusive to safety or ethics (author network), or how many papers are written by safety- or ethics-specific authors (paper network).

Bridge Concentration Fraction of Safety-Ethics shortest paths passing through a top-K set of authors (ranked by degree or centrality). Measures how concentrated cross-group brokerage is.

Average Shortest-Path (ASP) Median and mean shortest-path lengths for reachable pairs, both unweighted (hops) and weighted (distance = $\frac{1}{\texttt{weight}}$) . Computed for within-group (Safety-Safety, Ethics-Ethics) and cross-group (Safety-Ethics) pairs, with $\Delta = ASP_{S,E} - \texttt{mean}(ASP_{S,S}, ASP_{E,E})$ to measure relative separation. Weighted Average Shortest Path allows us to answer the question: for any two authors (or papers), how many papers-in-common (or authors-in-common) would it take to meet? This measures the distance of direct collaboration, providing a measure of the activity in research silos. We aim to demonstrate that aggregations of these ASPs show **more hops** and **greater distance** across disciplines.

3.2.2 Statistical Significance Testing

To ensure metrics reflect structural properties rather than artifacts, we employ multiple null models [26]:

Label-Shuffle Null Permute safety/ethics labels across fixed topology (500-2000 reps), recompute metrics, and compute empirical p-values and z-scores [46].

Degree-Preserving Rewire Null Randomize edges while preserving degree sequences, recompute metrics to test if topology alone explains separation [51].

These tests confirm that observed separation (e.g., high homophily, concentrated bridges, fragile reachability) is statistically significant and not due to random chance or degree distribution. Each null model was run with 1,000 permutations; p-values are estimated as the fraction of nulls exceeding observed values.

4 Results and Discussion

4.1 Homophily

Our analysis of **homophily**—the tendency for researchers to collaborate with others from their own community—provides the starkest evidence of the schism (See "Baseline" in Table 2). In our author network, we measured a global homophily of **83.1%**. This indicates that **over four out of every five collaborations** occur between authors who are exclusively focused on either safety or ethics, demonstrating a powerful in-group preference. This insularity holds when examining each community individually: **73.5%** of collaborations involving a safety researcher are with other safety specialists, while **68.2%** of collaborations for ethics researchers are with their peers. These findings are statistically significant under shuffle-based and degree-preserving rewiring null models ($p \ll 0.01$ for both).

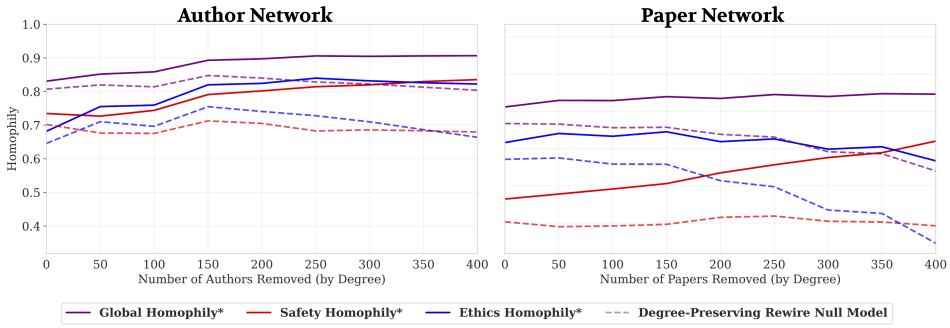

Figure 3: Author-based (left) and paper-based (right) networks exhibit high homophily globally and in safety/ethics networks. They increase as top authors are pruned ($p \ll 0.05$).

Table 2: Homophily values under degree-based removal of top authors

	Author Network			Paper Network		
# Authors Removed	Global	Safety	Ethics	Global	Safety	Ethics
Baseline (0)	83.1%	73.5%	68.2%	71.2%	46.5%	61.6%
100	85.8%	74.4%	75.9%	72.9%	49.2%	63.3%
200	89.7%	80.2%	82.4%	73.5%	53.5%	61.9%
300	90.5%	82.0%	83.2%	74.0%	57.6%	59.9%
400	90.7%	83.5%	82.2%	74.7%	62.0%	56.7%

Interestingly, the effect is slightly less pronounced in our paper-based network (homophily of **71.2%**), suggesting that while researchers themselves remain highly siloed, a small number of cross-disciplinary papers create connections that make the *literature* appear more integrated than the *social network* of its authors. This discrepancy highlights the outsized impact of a small number of **"bridge" authors**. A single interdisciplinary author, by co-authoring one paper with the "pure safety" community and another with the "pure ethics" community, creates a direct link between those two otherwise disconnected bodies of literature. This single author's work acts as a hub, connecting entire clusters of papers and significantly lowering the paper network's homophily. The result reveals a critical insight: while the social network of researchers remains highly segregated, the literature itself is stitched together by the crucial, yet sparse, work of these bridging individuals.

To demonstrate this, we conduct a perturbation test by removing the nodes with the highest degrees and observing the effect on homophily (See Figure 3). In doing so, we show how the loss of well-connected bridge authors impacts total homophily. The author-based network shows a persistent increase, rising to 90.7% global homophily. For papers, removing high-degree papers (papers written by prolific co-authors) contributes to a slight increase in homophily globally, but the effect is dispersed—while ethics research maintains its current rate with a slight decrease, the homophily of safety research *increases* by 15%. The divided impact suggests that the relative diversity of safety research is concentrated in "hub" papers—removing them quickly disconnects safety research from ethics-based or mixed-method literature.

4.2 Bridge Connectivity

The results in Table 3 reveal two critical structural properties of the safety-ethics network: **highly concentrated brokerage** and **extreme fragility**. At the baseline, we find that a small elite of authors—just the **top 1%** by degree—act as brokers for a disproportionate **58.0%** of all shortest paths between the two communities. This concentration is statistically significant ($p < 0.01$) and indicates that cross-disciplinary communication relies on a "hub-and-spoke" model rather than a broad, distributed dialogue.

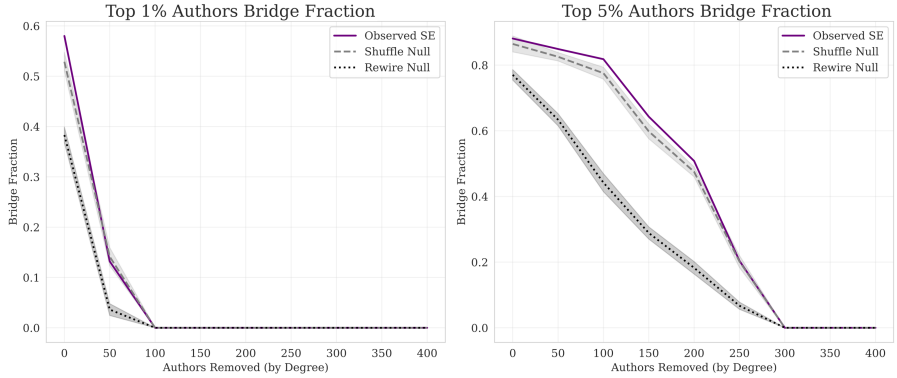

Figure 4: Author-based bridge connectivity with authors removed by degree. 0 is the baseline, with shuffle and degree-preserving nulls included (p « 0.05).

Furthermore, this connectivity is exceedingly fragile. The targeted removal of these high-degree authors causes a precipitous drop in connectivity that far outpaces the decay in our null models. After removing just 100 authors, for instance, the most central bridge paths (those controlled by the top 1%) vanish entirely (**0.0%**), and overall connectivity plummets. This demonstrates that the intellectual exchange between AI safety and ethics is not a robust, resilient conversation but a tenuous one, critically dependent on a few central individuals.

Table 3: Fraction of bridge paths passing through top authors, under degree-based removal

	Observed Network		Label-Shuffle Null		Degree-Rewire Null	
# Authors Removed	Top 1%	Top 5%	Top 1%	Top 5%	Top 1%	Top 5%
0 (Baseline)	**58.0%** (p<0.01)	**88.1%** (p<0.05)	52.9%	86.4%	38.3%	77.0%
50	**13.2%** (p>0.05)	**84.9%** (p<0.05)	14.1%	82.5%	3.6%	63.4%
100	**0.0%**	**81.8%** (p<0.01)	0.0%	77.5%	0.0%	44.2%
150	**0.0%**	**64.3%** (p<0.05)	0.0%	59.8%	0.0%	28.9%
200	**0.0%**	**50.8%** (p<0.05)	0.0%	47.5%	0.0%	18.3%

4.3 Weighted Average Shortest Path

Our analysis of the average shortest paths between authors reveals a subtle but persistent "friction" that inhibits cross-disciplinary collaboration. While the mean path lengths are superficially similar,

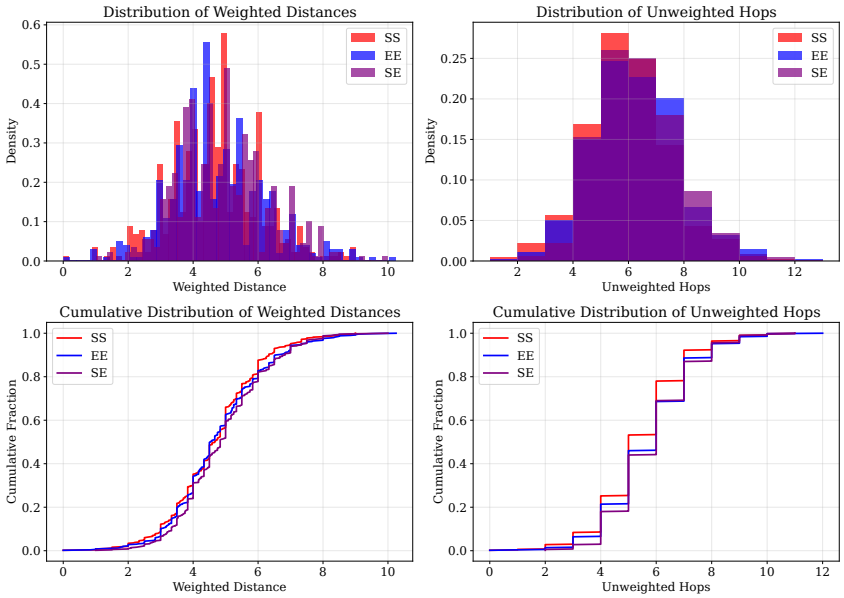

Figure 5: Distributions of weighted path distances (left) and unweighted hops (right) for pairs of authors. The Safety-Ethics (SE) distribution is systematically shifted to the right compared to within-group Safety-Safety (SS) and Ethics-Ethics (EE) pairs, indicating longer paths are required to connect researchers across the two communities.

their underlying distributions, shown in Figure 5, tell a clearer story. The cumulative distribution for cross-community (SE) paths is systematically shifted to the right, meaning a consistently longer weighted distance is required to connect a random pair of safety and ethics researchers compared to pairs within their own communities ($p < 0.01$). For instance, **80%** of all Safety-Safety author pairs can be connected within a weighted distance of 6.0, whereas reaching the same fraction of Safety-Ethics pairs requires a longer path.

This separation is further confirmed by our reachability analysis, which measures the fraction of author pairs connected within a given number of hops. We find that cross-group reachability is significantly lower than would be expected by chance. After five hops, only **16.9%** of safety-ethics author pairs are connected, a figure substantially below the **21.5%** we would expect in a randomly labeled network with the same structure ($p < 0.001$). This demonstrates that the path to collaboration is not only longer on average but also structurally impeded, quantitatively confirming the schism between the two fields.

5 Limitations

While this study provides the first large-scale, quantitative evidence of the schism between AI safety and ethics, we acknowledge several limitations that define the scope of our findings.

Scoping and Data Representation Our analysis is scoped to the proceedings of 12 major academic conferences from 2020-2025. While these venues are central to the field, our corpus does not capture the full research ecosystem, which includes important work published in journals, workshops, industry technical reports, and influential preprints. Our focus on the post-2020 era captures the modern, LLM-centric landscape but excludes foundational work that shaped the communities prior to this period. Furthermore, while we provide a robust justification for excluding arXiv preprints, this choice means our analysis may not capture the most recent or fast-moving research trends that have yet to undergo peer review.

Methodological Assumptions Our methodology relies on several key assumptions. First, our two-stage filtering process, while rigorously validated, necessarily operationalizes "safety" and "ethics" into discrete categories. These fields are complex and evolving, and a different concep-

tual taxonomy could yield different quantitative results. Second, we use co-authorship networks as a proxy for **active collaboration**. This is a strong signal of direct partnership but does not measure other forms of intellectual exchange, such as the passive influence captured by citation networks. Our findings are therefore a measure of social and institutional separation, not necessarily a complete lack of intellectual cross-pollination. Finally, our author network analysis focuses on researchers with at least two publications in our corpus. This standard technique allows for a robust analysis of the core research community but means our conclusions may be less representative of newcomers or authors with a single relevant publication.

6 Conclusion

The schism between AI safety and ethics is not merely anecdotal; it is a structural reality with measurable consequences. Our network analysis of over 6,000 papers provides quantitative evidence of this divide, revealing a landscape characterized by high homophily (**83.1%** in-group collaboration) and fragile, concentrated brokerage. This structural gap persists despite significant thematic overlap, suggesting that the barriers to integration are more social and institutional than they are intellectual.

The consequences of this fragmentation extend beyond academia into global policy, creating a fractured approach to governance where our most pressing risks are treated as competing priorities. Landmark international reports on AI safety, for instance, have focused heavily on technical control and existential risk while largely neglecting the critical issues of bias and fairness central to the ethics community [9]. Conversely, influential frameworks like UNESCO's recommendations on AI ethics provide robust normative guidance on fairness but offer fewer tractable, technical solutions for implementation and do not engage with long-term catastrophic risks [50]. Each community, operating in isolation, exports an incomplete vision of "alignment" to policymakers.

It is precisely this gap that a venue like IASEAI is designed to fill. Its mission explicitly combines the language of safety, ethics, and alignment, creating a unique intellectual space for the cross-pollination our findings show is urgently needed. Yet, the very structure of our field's calls for papers—including IASEAI's—still tends to separate these topics into distinct tracks, underscoring how deeply ingrained this divide has become. Our work is a call to move beyond these inherited categories. We argue that the future of alignment research depends not on choosing between control and justice, but on synthesizing them into a unified, resilient, and truly effective discipline. We conclude with three pragmatic recommendations to accomplish such a future:

1. Shared Empirical Benchmarks. The empirical standards of the two fields are almost entirely disjoint. Safety relies on adversarial red-teaming to probe for catastrophic failures, while ethics employs sociotechnical audits to uncover embedded biases [44]. We advocate for the creation of **shared evaluative benchmarks** that treat alignment not as a bifurcated concept but as a unified property. For instance, a benchmark could require a model to demonstrate robustness to jailbreaking while simultaneously satisfying group fairness constraints, forcing a direct confrontation with the safety-utility tradeoffs that currently divide the fields.

2. Cross-Institutional Venues. Our network analysis reveals a divide sustained by institutional geography—separate conferences, funding streams, and academic departments. To bridge this, our second pathway calls for **structural interventions in our academic venues**. We laud IASEAI for acting as an inaugural act of unification, but we advocate that the venue and its organizers lean into interdisciplinary thought—pioneering joint conference tracks, workshops, and doctoral consortia that require co-authorship between technical and social scientists. By creating spaces where collaboration is not just encouraged but expected—and, critically, evaluated by mixed review panels—we can lower the high social and professional cost of crossing the disciplinary divide our data reveals.

3. Integrative Research Methodologies. Finally, we propose a pathway of **methodological synthesis**. The tools of one community can solve the problems of the other. The mechanistic interpretability techniques developed in AI safety, for example, are powerful instruments for conducting the deep algorithmic audits called for by AI ethics. Conversely, the participatory design methods from ethics offer a robust framework for operationalizing the "human values" that scalable oversight in AI safety aims to align with. We advocate for research that explicitly couples these approaches as a concrete means of forging a unified, sociotechnical practice of alignment.

Acknowledgments and Disclosure of Funding

This project was independently supported by the authors, without external funding. We extend our gratitude to Lauren Klein, Daphne Ippolito, and Eva Roytburg for reviewing early versions of this manuscript.

References

[1] ACL anthology bibliography (2025 XML dump). https://aclanthology.org/anthology+abstracts.bib.gz, 2025. ACL snapshot used for author/venue metadata, accessed 2025-10-11.

[2] DBLP computer science bibliography (2025 XML dump). https://dblp.org/xml/, 2025. Data snapshot used for author/venue metadata, accessed 2025-10-11.

[3] Moses Alabi and Tobi Holmes. AI safety and ethics: Developing robust frameworks for ethical AI development and deployment. ResearchGate, 2024. URL https://www.researchgate.net/publication/383283895.

[4] Alignment Forum. Outer vs. inner misalignment: Three framings. https://www.alignmentforum.org/posts/poyshiMEhJsAuifKt/outer-vs-inner-misalignment-three-framings-1, 2022. Accessed: 2025-10-10.

[5] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete Problems in AI Safety, July 2016. URL https://arxiv.org/abs/1606.06565. arXiv:1606.06565 [cs].

[6] Anthropic. HH-RLHF: Helpfulness and harmlessness RLHF dataset. https://github.com/anthropics/hh-rlhf, 2022. Accessed: 2025-10-10.

[7] Michael C Behrent. Foucault and technology. *History and Technology*, 29(1):54–104, 2013.

[8] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623, 2021.

[9] Yoshua Bengio, Sören Mindermann, Daniel Privitera, et al. International AI Safety Report, January 2025. URL https://arxiv.org/abs/2501.17805. arXiv:2501.17805 [cs].

[10] Ruha Benjamin. Race after technology. In *Social Theory Re-Wired*, pages 405–415. Routledge, 2023.

[11] Leonard F. Bereska and Efstratios Gavves. Mechanistic interpretability for AI safety — A review. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=ePUVetPKu6.

[12] Jason Borenstein, Frances S. Grodzinsky, Ayanna Howard, Keith W. Miller, and Marty J. Wolf. AI Ethics: A Long History and a Recent Burst of Attention. *Computer*, 54(01):96–102, January 2021. ISSN 0018-9162. doi: 10.1109/MC.2020.3034950. URL https://www.computer.org/csdl/magazine/co/2021/01/09321834/1qmbkXCazy8.

[13] N. Bostrom. *Superintelligence: Paths, Dangers, Strategies*. OUP Oxford, 2014. ISBN 978-0-19-166682-7.

[14] Nick Bostrom. What happens when our computers get smarter than we are? TED Talk, 2015. URL https://www.ted.com/talks/nick_bostrom_what_happens_when_our_computers_get_smarter_than_we_are.

[15] Brookings Institution. With AI, we need both competition and safety. https://www.brookings.edu/articles/with-ai-we-need-both-competition-and-safety/, 2024. Accessed: 2025-10-10.

[16] Joy Buolamwini. *Unmasking AI: My Mission to Protect What Is Human in a World of Machines*. Random House, 2023.

[17] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, pages 77–91. PMLR, 2018.

[18] Ben Chester Cheong. Transparency and accountability in AI systems: Safeguarding wellbeing in the age of algorithmic decision-making. *Frontiers in Human Dynamics*, 6:1421273, 2024. doi: 10.3389/fhumd.2024.1421273.

[19] Luca Deck, Jakob Schoeffer, Maria De-Arteaga, and Niklas Kühl. A critical survey on fairness benefits of explainable AI. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1579–1595. ACM, 2024. doi: 10.1145/3630106.3658990.

[20] Catherine D'Ignazio and Lauren F Klein. *Data Feminism*. MIT Press, 2023.

[21] Murat Durmus. The difference between AI safety, AI ethics, and responsible AI. Medium, 2024. URL https://murat-durmus.medium.com/the-difference-between-ai-safety-ai-ethics-and-responsible-ai-8296306af427.

[22] Jonathan Vanian Elias and Hayden Field Jennifer. AI research takes a backseat to profits as Silicon Valley prioritizes products over safety, experts say, May 2025. URL https://www.cnbc.com/2025/05/14/meta-google-openai-artificial-intelligence-safety.html.

[23] Timnit Gebru. Effective Altruism Is Pushing a Dangerous Brand of 'AI Safety'. *Wired*, December 2022. URL https://www.wired.com/story/effective-altruism-artificial-intelligence-sam-bankman-fried/.

[24] Google Research. VaultGemma: The world's most capable differentially private LLM. https://research.google/blog/vaultgemma-the-worlds-most-capable-differentially-private-llm/, 2024. Accessed: 2025-10-10.

[25] Ryan Greenblatt, Buck Shlegeris, Kshitij Sachan, and Fabien Roger. AI control: Improving safety despite intentional subversion. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 16295–16336. PMLR, 2024. URL https://proceedings.mlr.press/v235/greenblatt24a.html.

[26] Milad Haghani. What makes an informative and publication-worthy scientometric analysis of literature: A guide for authors, reviewers and editors. *Transportation Research Interdisciplinary Perspectives*, 22:100956, November 2023. ISSN 2590-1982. doi: 10.1016/j.trip.2023.100956. URL https://www.sciencedirect.com/science/article/pii/S2590198223002038.

[27] Hugo Horta, Shihui Feng, and João M. Santos. Homophily in higher education research: a perspective based on co-authorships. *Scientometrics*, 127(1):523–543, January 2022. ISSN 1588-2861. doi: 10.1007/s11192-021-04227-z. URL https://doi.org/10.1007/s11192-021-04227-z.

[28] Brian Judge, Mark Nitzberg, and Stuart Russell. When code isn't law: rethinking regulation for artificial intelligence. *Policy and Society*, 44(1):85–97, April 2025. ISSN 1449-4035. doi: 10.1093/polsoc/puae020. URL https://doi.org/10.1093/polsoc/puae020.

[29] Kazi Zainab Khanam, Gautam Srivastava, and Vijay Mago. The homophily principle in social network analysis: A survey. *Multimedia Tools and Applications*, 82(6):8811–8854, 2023.

[30] Ang Li, Yichuan Mo, Mingjie Li, Yifei Wang, and Yisen Wang. Are smarter llms safer? exploring safety-reasoning trade-offs in prompting and fine-tuning, 2025. URL https://arxiv.org/abs/2502.09673.

[31] Zhiqiang Lin, Huan Sun, and Ness Shroff. AI Safety vs. AI Security: Demystifying the Distinction and Boundaries, June 2025. URL https://arxiv.org/abs/2506.18932. arXiv:2506.18932 [cs].

[32] Sasha Luccioni. AI is dangerous, but not for the reasons you think. TED Talk, 2023. URL https://www.ted.com/talks/sasha_luccioni_ai_is_dangerous_but_not_for_the_reasons_you_think.

[33] Mario Malički, Maria Janina Sarol, and Juan Pablo Alperin. Analyzing preprints: The challenges of working with metadata from arXiv's Quantitative Biology section. Scholarly Communications Lab, November 2019. URL https://www.scholcommlab.ca/2019/11/07/preprints-challenges-part-four/.

[34] McKinsey & Company. As gen AI advances, regulators and risk functions rush to keep pace. https://www.mckinsey.com/capabilities/risk-and-resilience/our-insights/as-gen-ai-advances-regulators-and-risk-functions-rush-to-keep-pace, 2025. Accessed: 2025-10-10.

[35] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6):1–35, 2021. doi: 10.1145/3457607.

[36] Bahar Memarian and Tenzin Doleck. Fairness, accountability, transparency, and ethics (FATE) in artificial intelligence (AI) and higher education: A systematic review. *Computers and Education: Artificial Intelligence*, 5:100152, 2023. doi: 10.1016/j.caeai.2023.100152.

[37] Edoardo Mosca, Ferenc Szigeti, Stella Tragianni, Daniel Gallagher, and Georg Groh. SHAP-based explanation methods: a review for NLP interpretability. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4593–4603, 2022.

[38] Luke Munn. The uselessness of ai ethics. *AI and Ethics*, 3(3):869–877, 2023.

[39] Rajakishore Nath and Riya Manna. From posthumanism to ethics of artificial intelligence. *AI & SOCIETY*, 38(1):185–196, 2023.

[40] Mark EJ Newman. Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences*, 101(suppl_1):5200–5205, 2004.

[41] Mark EJ Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69(2):026113, 2004.

[42] Claudio Novelli, Mariarosaria Taddeo, and Luciano Floridi. Accountability in artificial intelligence: What it is and how it works. *AI & Society*, 39(4):1871–1882, 2024.

[43] Hammond Pearce, Baleegh Ahmad, Benjamin Tan, Brendan Dolan-Gavitt, and Ramesh Karri. Asleep at the keyboard? Assessing the security of GitHub Copilot's code contributions. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 754–768. IEEE, 2022. doi: 10.1109/SP46214.2022.9833571.

[44] Richard Ren, Steven Basart, Adam Khoja, Alice Gatti, Long Phan, Xuwang Yin, Mantas Mazeika, Alexander Pan, Gabriel Mukobi, Ryan H. Kim, Stephen Fitz, and Dan Hendrycks. Safetywashing: Do AI safety benchmarks actually measure safety progress? In *Advances in Neural Information Processing Systems*, volume 37. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/7ebcdd0de471c027e67a11959c666d74-Abstract-Datasets_and_Benchmarks_Track.html.

[45] Malak Sadek, Emma Kallina, Thomas Bohné, Céline Mougenot, Rafael A Calvo, and Stephen Cave. Challenges of responsible ai in practice: scoping review and recommended actions. *AI & Society*, 40(1):199–215, 2025.

[46] Ayushi Saxena. *The Shuffling Effect: Vertex Label Error's Impact on Hypothesis Testing, Classification, and Clustering in Graph Data*. PhD thesis, University of Maryland, College Park, 2024.

[47] Guobin Shen, Dongcheng Zhao, Yiting Dong, Xiang He, and Yi Zeng. Jailbreak antidote: Runtime safety-utility balance via sparse representation adjustment in large language models. In *Proceedings of the 13th International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=s20W12XTF8.

[48] Torben Swoboda, Risto Uuk, Lode Lauwaert, Andrew P. Rebera, Ann-Katrien Oimann, Bartlomiej Chomanski, and Carina Prunkl. Examining Popular Arguments Against AI Existential Risk: A Philosophical Analysis, January 2025. URL https://arxiv.org/abs/2501.04064. arXiv:2501.04064 [cs].

[49] Iman Tahamtan and Lutz Bornmann. What do citation counts measure? An updated review of studies on citations in scientific documents published between 2006 and 2018. *Scientometrics*, 121(3):1635–1684, 2019. doi: 10.1007/s11192-019-03243-4.

[50] UNESCO. Recommendation on the ethics of artificial intelligence. Technical report, UNESCO, 2021. URL https://www.unesco.org/en/artificial-intelligence/recommendation-ethics.

[51] František Váša and Bratislav Mišić. Null models in network neuroscience. *Nature Reviews Neuroscience*, 23(8):493–504, 2022.

[52] Sahil Verma and Julia Rubin. Fairness definitions explained. In *Proceedings of the 2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*, pages 1–7. ACM, 2018. doi: 10.1145/3194770.3194776.

[53] Anvesh Rao Vijjini, Somnath Basu Roy Chowdhury, and Snigdha Chaturvedi. Exploring safety-utility trade-offs in personalized language models. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11316–11340, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.565. URL https://aclanthology.org/2025.naacl-long.565/.

[54] E. Yudkowsky and N. Soares. *If Anyone Builds It, Everyone Dies: Why Superhuman AI Would Kill Us All*. Little, Brown, 2025. ISBN 978-0-316-59566-7.