The Double-Edged Sword: A Framework for Analyzing and Forecasting Capability Spillovers from AI Safety Research

Contributors: Ihor Kendiukhov, Santi Bent, Serena Ives, Hemang Jain, Varvara Lantukh, Alexandra Moraru, Clinton Morimoto, Aditya Raj, Nathan Theng, Troy Tian, Chris Underhill

Emails: kendiukhov@gmail.com, santi1bent@gmail.com, ivesserena@gmail.com, hemang.jain@students.iiit.ac.in, varlantukh@gmail.com, am1563@student.london.ac.uk, morimotoclinton@gmail.com, rajsecrets03@gmail.com, thengnathan@gmail.com, troyhtian@gmail.com, chrispunderhill@gmail.com

Abstract

AI safety research routinely yields techniques that, once published, accelerate general AI capabilities, with famous canonical examples like reinforcement learning from human feedback (RLHF). Yet the field has no systematic method for anticipating this dynamic before research is funded or released. This report presents the Capability Spillover Potential (CSP) framework: a multidimensional scoring rubric that estimates the risk of a safety research direction being co-opted for capability advancement. We first ground the framework in retrospective case studies, including RLHF, mechanistic interpretability, and process reward models. We formalize it as a behaviorally anchored rubric scored on ordinal scales and combine two independent data streams (expert survey responses and in-house scoring) through an explicit aggregation function. Applied across twelve core AI safety areas, the framework produces interpretable CSP estimates, with RL-based alignment scoring highest. We argue that capability spillover is a structural, recurring, and scorable phenomenon and that making it legible enables funders, researchers, and governance bodies to pursue differential progress more deliberately. We are explicit about the framework's current limitations, particularly the assumption that spillover risk is stable enough to forecast prospectively.

Introduction

The case for AI safety research rests on a simple premise: powerful AI systems pose risks that must be managed before, not after, those systems are deployed (Amodei et al., 2016). But safety research does not occur in a vacuum. It is produced inside a competitive ecosystem in which investment in raw capability vastly outstrips investment in safety and in which any technique that confers a performance advantage is rapidly identified, published, and absorbed into capability scaling efforts, regardless of the intentions of the researchers who produced it. The result is a feedback loop that is uncomfortable for the field to confront. Research intended to make AI safer can supply precisely the insights needed to make it more powerful.

This is not a hypothetical worry, either. RLHF, for instance, was developed to align language models with human preferences, yet it became a core technical driver behind ChatGPT and the industry wide assistant race that followed. Mechanistic interpretability, conceived as “neuroscience for AI” to audit models for deception, has produced tools that also enable fine-grained behavioral control. Process reward models, rooted in scalable oversight proposals, contributed directly to the reasoning model paradigm. In each case the spillover was not a misuse of the work but a direct consequence of the knowledge it created.

The problem this report addresses is that the safety community has no systematic way to anticipate such spillovers. Research directions are chosen, funded, and published largely without formal analysis of their capability externalities. Our thesis is that capability spillover is a structural and recurring phenomenon, that it operates through identifiable mechanisms, and that it can therefore be made legible and scorable. We develop this claim in three steps: we show that spillover recurs across distinct safety areas through common mechanisms, we present the CSP framework that operationalizes these mechanisms into a scoring rubric and aggregation method, and we report preliminary results across twelve safety areas while being explicit about where the argument remains unproven.

Background & Setup

Our work builds on the concept of *differential progress* (commonly known as *differential technological development*), the principle, articulated in earlier AI risk scholarship, that the safety-relevant capabilities of a technology should advance faster than its general or risk-increasing capabilities (Bostrom, 2014). Under this principle, a safety research program is not automatically beneficial; its net value depends on the *margin* by which it advances safety over capabilities. A technique that improves safety somewhat but capabilities substantially can be net negative for differential progress even though it was produced with safety intentions.

We define two terms precisely. *Capability spillover* is the process by which safety-motivated research produces insights, tools, or methods that are subsequently repurposed to advance general capabilities. *Capability Spillover Potential* (CSP) is our proposed metric for estimating, in advance, how susceptible a given research direction is to such repurposing.

The framework is explicitly modeled on dual-use risk assessment in biosecurity, where evaluating whether research could be co-opted for harm is standard practice and is institutionalized in funding and publication review. AI safety has no equivalent methodology. Existing discussions of capability spillover within the field are real but largely informal, anecdotal, and retrospective. They observe that RLHF “leaked” into

capabilities after the fact, without ever having offered a structured way to have anticipated it.

Methodologically, our approach is *meta-research*: we treat AI safety research itself as the object of study. The CSP framework is designed to operate at two levels of granularity — individual papers and whole research areas, using structured rubrics with behavioral anchors scored on ordinal scales by independent reviewers. This report presents the framework in the state reached at the end of the SPAR project cycle, including its current rubric version, its dual-source data architecture, and a first application across twelve safety areas.

Main Argument

Spillover is structural, not incidental

Our first claim is that capability spillover is not a series of unlucky accidents but a structural feature of how safety research interacts with a capability hungry ecosystem. The evidence is the recurrence of identifiable spillover *mechanisms* across independent research areas. RLHF, building on deep reinforcement learning from human preferences (Christiano et al., 2017), made language models steerable, converting them from text completers into commercially viable assistants and triggering an industry wide capabilities race (Ouyang et al., 2022); the mechanism here is that an alignment technique unlocked an entire product category. Mechanistic interpretability produced capability improving methods. For example, the finding that refusal behavior is mediated by a single direction in activation space, that are equally usable for safety monitoring and for fine-grained capability or behavioral control (Olah et al., 2020; Arditi et al., 2024); the mechanism here is that diagnostic understanding is intrinsically also a control lever. Process reward models, rooted in scalable oversight proposals (Christiano et al., 2018; Irving et al., 2018), contributed to the emergence of reasoning models and new test time scaling laws (Lightman et al., 2023; Snell et al., 2024); the mechanism here is that a better oversight signal is also a better training signal. Three independent areas, three distinct mechanisms, one pattern: the knowledge that makes a model safer to use frequently also makes it more capable. Because the mechanisms recur, they can be enumerated, and what can be enumerated can, in principle, be scored prospectively.

From mechanisms to a rubric

The CSP framework operationalizes these mechanisms as scoring dimensions. Our most developed version, the V1.2 rubric, scores a research area on four dimensions: *Direct Capability Impact* (does the work directly improve performance on capability benchmarks?), *Technical Transferability* (how easily do its outputs transfer into capability applications?), *Audience and Venue Exposure* (how visible is the work to capability researchers?), and *Strategic Leverage* (does it open new capability research

directions?). Each dimension is scored on an ordinal scale with behavioral anchors, so that scorers are calibrated against concrete descriptions rather than unaided intuition.

CSP Rubric V1.2

Direct Capability Impact

Does the paper itself demonstrate or report improvements on recognised capability benchmarks (e.g., MMLU, HumanEval, GSM8K, ARC, BigBench, SWE-Bench, HellaSwag, etc.)? This dimension captures whether the paper's own experiments already show capability gains, regardless of whether that was the authors' intent.

Component Description	Rating	Uncertainty
Capabilities Benchmarks 1pt - Does it use any benchmarks to validate research that aren't purely safety-based? 1pt - Either: Does the work rely on many (>3) capabilities benchmarks or use any well-recognized capabilities benchmarks in the field? *e.g. MMLU, HumanEval, SWE-Bench*		
Capabilities Gains 1pt - Does the paper achieve at least minor incidental gains (1-3% relative improvement)? 1pt - Does it see substantial gains, in excess of 5% relative improvement? 1pt - Does it achieve state-of-the-art or nearly leading capabilities results - giving a publishable result regardless of any safety training?		
Total - 5pts		

Audience & Venue Exposure

How visible is this work to capabilities researchers? Considers publication venue, author affiliations and networks, framing and terminology, and observed citation patterns.

Component Description	Rating	Uncertainty
Venue Exposure 1pt - Was this work published in a "top-tier" venue for capabilities research or general AI / ML publications? *e.g. NeurIPS, ICML, ICLR* 1pt - Has this research - including its results or methods - received widespread press coverage or otherwise appeared in highly-visible settings?		
Author Background 1pt - Have a sizable percentage of the work's authors specialized primarily in capabilities research in the recent past? 1pt - Is this work itself or a sizable percentage of the work's authors closely associated with any major frontier labs? *e.g. OpenAI, Anthropic, Google DeepMind*		
Target Audience 1pt - Is the work generally framed, at least in major part, as a capabilities piece, versus a purely safety-oriented piece?		
Total - 5pts		

Technical Transferability

Does the paper produce concrete outputs (e.g. architectures, training procedures, datasets, benchmarks, algorithms) that a capabilities researcher could directly extract and apply to improve model performance on general tasks? The more directly usable the output, the higher the score.

Component Description	Rating	Uncertainty
Extractable Outputs 1pt - Are there any extractable research outputs adjacent to capabilities on a technical level? *Exceptions include purely conceptual or governance-focused approaches.* 1pt - Does the paper reveal any insights about model behavior or additional relevant parameters that may inform capabilities research?		
Transferable Models 1pt - Does the paper introduce an approach directly applicable to capabilities work with at most minor modification? *e.g. a data-filtering pipeline or a training efficacy improvement* 1pt - Could a capabilities researcher adopt core models or algorithms as-is without modification?		
Transferable Assets 1pt - Does the work create transferable assets to capabilities research, like large curated datasets or readily-accessible methods for pruning data?		
Total - 5pts		

Strategic Leverage

Does this paper open specific, identifiable new research directions that could lead to capability advances? Assesses whether the paper provides tools, frameworks, theoretical insights, or empirical findings that would enable concrete new lines of capability research that did not previously exist.

Component Description	Rating	Uncertainty
Marginal Contribution 1pt - Does the paper introduce a research direction or branch which has potential for capabilities researchers? 1pt - Are the paper's methods or findings theoretically valuable for inspiring or motivating capabilities research?		
Enablement 1pt - Is the work able to substantially lower the barrier to entry for a class of capability experiments? 1pt - Does the paper establish an entirely new paradigm within both its own safety field and within broader capabilities research? 1pt - Does the work spawn a benchmark suite or foundational method which enables capabilities research? *e.g. a way of eliciting latent model capabilities that generalizes broadly*		
Total - 5pts		

Figure 1: The CSP Rubric V1.2, used as described above.

Two data streams and an aggregation function

Estimating CSP at the level of a whole research area requires more input than any single scorer can provide, so the framework draws on two deliberately independent data streams. The first is *public response*: a publicly released survey targeted at three respondent tiers. Prominent figures, researchers at frontier labs, and members of independent or small organizations, together with data from a public website that lets anyone apply the framework. The second is *in-house scoring*: structured scoring by the

project team, more research driven and finer-grained, with an inter-rater reliability protocol that flags areas where two scorers diverge sharply so the divergence can be reconciled.

The following outlines our mathematical modeling procedure for using the survey-based data. We index each survey participant with an index i for the sake of notation. We index each of the twelve core research areas with an index j, following the previously established order. A scale check has survey participants scale three vignettes on one of our rubric's dimensions. Call the scores x_i, y_i, and z_i. Then, the participants choose one of those vignettes to give a lower-bound score l_i and upper-bound score r_i such that the participant ascertains an approximately eighty-percent chance of the "true" value lying between l_i and r_i.

Participants score their familiarity in each research area, from $0-3$, which we'll note as $v_{i,j}$. For each of the twelve research areas, participants assign scores from $0-5$ and confidence from $0-1$ in each of 5 dimensions, indexed k. Denote each score as $s_{i,j,k}$ and confidence as $c_{i,j,k}$. Finally, participants score each of the twelve research areas on safety benefit scores (its potential to improve AI safety) and its associated confidence; we denote these as $b_{i,j}$ and $o_{i,j}$ respectively.

Next, we detail our notation for the in-house scoring method. Using the V1.2 rubric, backing up scoring with rationale and specific research papers, we score these twelve research areas with scores of $0-5$ and uncertainty from $0-1$ (note the use of 'uncertainty' versus 'confidence'). More precisely than our survey data, we allow both our scores and our uncertainty to have step sizes of 0.1. If we label each in-house scorer with index i, each research area with index j, and each of the four (note four instead of five dimensions in this iteration of the rubric) with index k, we denote each score as $t_{i,j,k}$ and uncertainty as $u_{i,j,k}$.

Then, we move onto aggregation. We start by defining variables based on calibration data for the i-th survey respondent.

$$\alpha_i = 1 - \frac{2(x_i+y_i+z_i)}{15}$$
$$\beta_i = 1 - \frac{2(r-l)}{5}$$

These terms are responsible for creating a counterweight, e.g. if a respondent is prone to giving high ratings, we naturally adjust values closer to a middle value of 2.5.

Adjusting our calibration variables with an arbitrary factor of $\frac{1}{5}$, we determine survey based CSP for safety area indexed j as

$$S_j = \frac{1}{n}\sum_{i=1}^{n}[(1+\frac{\alpha_i+\beta_i}{5})\frac{v_{i,j}}{3}\sum_{k=1}^{5}[\frac{s_{i,j,k}c_{i,j,k}}{25}]]$$

To clarify its meaning, we break down this formula into each step. We normalize $v_{i,j}$ and $s_{i,j,k}$ into the range $0-1$ by multiplying each by a factor of $\frac{1}{3}$ and $\frac{1}{5}$ respectively. We average out summations using $\frac{1}{n}$ and $\frac{1}{5}$ terms - the latter is combined with the previous normalization to give $\frac{1}{25}$. The mixed alpha and beta term is meant to balance out a respondent's overconfidence/underconfidence and tendency to overrate or underrate; note that this implies R_j isn't necessarily contained within the interval $0-1$, but instead can exceed 1 (although unlikely).

Optionally, we can factor in safety scores as

$$S'_j = \frac{1}{n}\sum_{i=1}^{n}[(1+\frac{\alpha_i+\beta_i}{5})(\frac{v_{i,j}}{3}\sum_{k=1}^{5}[\frac{s_{i,j,k}c_{i,j,k}}{25}]-\frac{b_{i,j}o_{i,j}}{5})]$$

placing it within an expected interval (but not guaranteed) of -1 to 1 Aggregating our in-house scoring system, we calculate CSP as

$$T_j = \frac{1}{2n}\sum_{i=1}^{n}[(1+\frac{\alpha_i+\beta_i}{5})\frac{v_{i,j}}{3}\sum_{k=1}^{5}[\frac{s_{i,j,k}c_{i,j,k}}{25}]]+\frac{1}{2m}\sum_{i=1}^{m}[\sum_{k=1}^{4}[\frac{t_{i,j,k}u_{i,j,k}}{20}]]$$

Preliminary results

Applied across twelve core AI safety areas, the framework produces a rank-ordered set of raw CSP scores spanning roughly 0.086 to 0.389, with RL-based alignment scoring highest. This is a reassuring sanity check rather than a surprise: the area most associated with the project's anchor case study (RLHF) is also the area the framework independently flags as highest risk, which is weak but genuine evidence that the rubric tracks something real. A companion view adjusts raw CSP by an estimate of safety benefit, because a high CSP score is not by itself a verdict against a research area. A technique with large spillover potential can still be worth pursuing if its safety benefit is larger still. This raw-versus-adjusted distinction is central to the framework's intended use and to avoiding the misreadings discussed below.

Figure 2: A plot of CSP scores for each of twelve core AI safety areas.

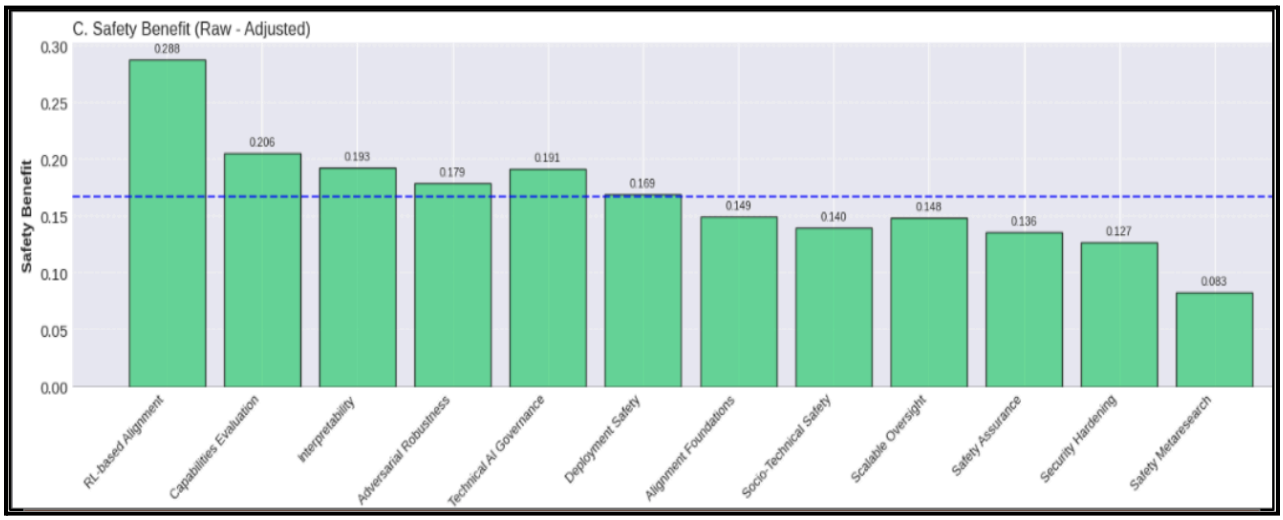

Figure 3: Aggregated safety benefit scores, in the order of Figure 1.

Where the argument is weakest

We flag three central assumptions, in keeping with the project's commitment to honest self-assessment. First, the framework's prospective value depends on spillover risk being stable enough to forecast, meaning that an area's CSP estimated today predicts spillover that materializes later. Our current evidence is retrospective; the planned expert survey is the main test of prospective stability, and until it is complete this assumption is unvalidated. Second, the framework assumes independent reviewers can apply the rubric consistently; without acceptable inter-rater reliability the scores are not trustworthy, and our reliability protocol, whether ultimately run as a full study or a two-person pilot, is still maturing. Third, the survey aggregation function reflects methodological choices that are reasonable but not the only defensible ones, and

several of its terms are still being finalized. Alternative choices could change the results, so the scores should be treated as provisional rather than settled.

Implications & Future Directions

If the central claim holds that capability spillover is structural, recurring, and scorable, several things follow for AI safety practice. Funders could incorporate CSP estimates into grant review alongside expected safety benefit, weighting the two explicitly rather than treating any safety labeled proposal as automatically net positive. Researchers could self-assess spillover risk before publication using a standardized checklist, making deliberate choices about framing, level of detail, and disclosure. Governance bodies could use CSP assessments to inform responsible disclosure norms and staged funding conditions, importing into AI the kind of dual-use review already normal in biosecurity.

These benefits are contingent, and the framework carries real downside risks that must be managed. A detailed map of how safety research becomes capability research is itself a roadmap that bad actors could exploit; CSP scores could be misread as funding verdicts and chill valuable but inherently dual-use work, with interpretability the obvious casualty; and the framework could be politicized to defund research groups on grounds other than scientific merit. The appropriate mitigations are to keep public outputs at the level of the framework and high-level ratings rather than how-to detail, to communicate consistently that CSP is one input among many and that raw CSP is explicitly not a recommendation, and to develop the framework openly so its biases can be surfaced and criticized.

The priorities for further validation are clear. Completing the expert survey would test prospective stability and external calibration; establishing inter-rater reliability would test scorer consistency; and fully documenting the aggregation function's term semantics would let others scrutinize and reproduce the scores. The framework should be regarded as a promising and usable first version, not a settled instrument.

Conclusion

Capability spillover from AI safety research is already happening, and it is currently happening unanalyzed. Our central claim is that this is not a run of bad luck but a structural dynamic that operates through a small set of recurring mechanisms, and that, because those mechanisms recur, the risk can be made legible and scored in advance. The CSP framework is our attempt to do so: a behaviorally anchored rubric, two independent data streams, an explicit aggregation function, and a first application across twelve core safety areas in which the highest-risk area is the one the project's anchor case study would predict.

The framework is not yet validated as a prospective instrument, and we have been explicit about the assumptions it rests on. But the alternative to an imperfect, improvable framework is the status quo: a field that continues to fly blind on a dynamic capable of neutralizing its own progress. Making capability spillover scorable is a precondition for pursuing differential progress deliberately rather than merely hoping for it.

References

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. *arXiv*. https://arxiv.org/abs/1606.06565

Arditi, A., Obeso, O., Syed, A., Paleka, D., Panickssery, N., Gurnee, W., & Nanda, N. (2024). Refusal in language models is mediated by a single direction. *Advances in Neural Information Processing Systems*.

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.

Christiano, P., Shlegeris, B., & Amodei, D. (2018). Supervising strong learners by amplifying weak experts. *arXiv*. https://arxiv.org/abs/1810.08575

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems*.

Irving, G., Christiano, P., & Amodei, D. (2018). AI safety via debate. *arXiv*. https://arxiv.org/abs/1805.00899

Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., Leike, J., Schulman, J., Sutskever, I., & Cobbe, K. (2023). Let's verify step by step. *arXiv*. https://arxiv.org/abs/2305.20050

Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., & Carter, S. (2020). Zoom in: An introduction to circuits. *Distill*. https://distill.pub/2020/circuits/zoom-in/

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*.

Snell, C., Lee, J., Xu, K., & Kumar, A. (2024). Scaling LLM test-time compute optimally can be more effective than scaling model parameters. *arXiv*. https://arxiv.org/abs/2408.03314