I work at the intersection of AI & Society, Communication and Discourse, with a focus on computational journalism and planning. My research aims to improve Generative AI and align it more closely with human values by studying the life cycle of human communicative acts: (1) Before (2) During, and (3) After communication.

- **Before: Modeling and Replicating Human Planning**: Current generative lack the ability to *plan* their outputs, especially for complex tasks with under-specified goals (e.g. structuring narratives [15, 18, 20, 13] or combining multiple informational sources [12, 19]). Since human plans are mostly *unobservable*, supervised data is scarce. To address this, I develop innovative methods to infer human planning strategies *at scale* using *data-driven approaches* [9, 12, 22, 23, 19, 15], which I use to train explicit planners.
- **During: Plan Aware Generation**: Even when provided with explicit prompts, generative models may deviate from intended objectives. This poses a particular challenge when we incorporate more explicit plans into generation. I have created techniques that help LLMs adhere more closely to complex prompts [11, 10, 16], reducing hallucinations and keeping the generated content aligned with user intentions.
- **After: Understanding AI's Societal Impact**: Once content is disseminated, its impacts on society need to be evaluated. In collaboration with Eric Horvitz at Microsoft Research, I demonstrated how misinformation propagates through search systems, leading to significant updates to Bing [14]. Additionally, I've examined how algorithmic decisions disproportionately affect certain demographics and proposed mitigation strategies [21]. Such methods are only increasing in importance the more we use algorithms to make decisions in our society. [21].

As a former New York Times journalist, I saw firsthand how high-quality information is found, confirmed and conveyed, and the critical role it plays in a healthy society. **My research vision, inspired by these experiences, is to build the next generation of AI systems to be more effective collaborators by enabling them to more deeply understand human processes**, giving us the tools to improve our informational ecosystem at scale.

My work has been supported by a 4-year Bloomberg PhD Fellowship. This has given me freedom to break ground while staying connected to journalists who could critique and improve my work. It is currently being used at Bloomberg, the *New York Times*, and Stanford Big Local News, impacting thousands of journalists. My methods have also been incorporated into major open-source code-bases: Huggingface, EleutherAI and IBM360. I have received recognition through Best and Outstanding Paper Awards (two (2) at EMNLP 2024 [13, 20], Computation+Journalism 2023 [8], NAACL 2022 [15]), Spotlight awards (ICML, 2024 [16]), and oral presentations (NAACL 2024 [18]).

I take pride in the growing community of researchers adopting similar approaches. I have collaborated with peers in areas such as: scientific discovery, video script editing, creative writing, musical composition and legal writing. I designed and successfully proposed an upcoming tutorial at NAACL 2025 to unify this field, which we call "**Creative Planning**". *I am also consulting with a team at OpenAI to continue to incorporate these ideas in o1.*

Finally, my work is having an impact in classrooms. I am partnering with USC Annenberg journalism professors to run classroom experiments testing how much journalism students can benefit from my research. My techniques have also been incorporated into courses at Stanford, USC, UCSD, Berkeley, and Columbia. My work has gained substantial media attention, including by Wired and the New York Times.

Before: Conceptualizing and Planning

For models to be effective machine-in-the-loop partners with humans in many complex tasks (e.g. in journalism [12], legal writing [18], etc.), they must better understand, then integrate, into human action sequences. For e.g., a journalist often follows multiple steps before drafting a news article (e.g. in Figure 1: "*develop angle*" → "*find sources*" → "*confirm facts*").

The first part of my work has laid the groundwork for *studying human actions in creative tasks*. Focusing on journalism, I have collaborated with professional journalists to study how humans conduct the news-gathering process, from: how stories are chosen [9], how sources are selected [12, 19, 13, 2], and how edits are made [15, 1]. These actions are often invisible in

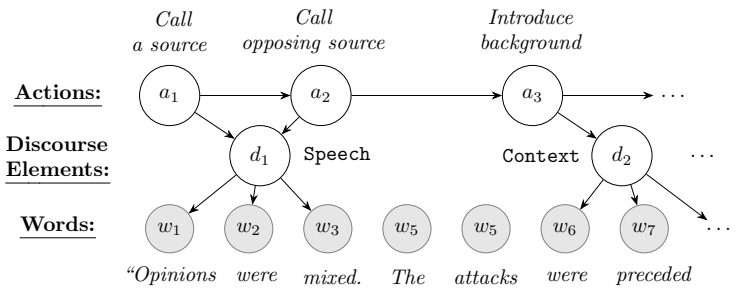

Fig. 1: A human's internal/external actions (i.e. a_1, a_2..) are inferrable via the text they write (w_1, w_2,...), expressed through (un)intentional discourse acts (d_1, d_2...). My primary research focus has been: (a) to show this, and show how this lens allows us to explore human creative processes (b) developing ways to imbue this into language models.

the final written piece; however, I have shown that they can be inferred with data-driven, unsupervised processes (**Outstanding Paper Award, EMNLP 2024**) [13, 20]. Specifically, the fundamental insight I have made is that we can extend discourse theory to show that certain cues in written text signal actions (or intentions) made by humans while writing, and these can be inferred. This is not surprising: domain experts can also "read between the lines" to deduce the process behind a peer's work. For e.g., when we read a scientific paper, as researchers, we can likely infer details like implementation choices, negative trials, or hyperparameter tuning, even if they are not explicitly stated.

To infer actions, I've developed a variety of methods. I first designed Bayesian models to learn clusters of co-occurring words that indicate *latent actions* [4, 7]. I manually annotated datasets to train classifiers [9, 12, 18, 19, 10, 1]. In more recent work, I developed novel approaches using large language models (LLMs) to identify actions, yielding unsupervised, end-to-end schemas *preferred even over expert-created schemas* [2, 3]. Finally, I showed that increasing visibility into intermediate states (i.e. writing drafts), can help us learn even more nuanced plans (**Outstanding Paper Award, NAACL 2022**) [15]. A second question emerged: since plans are latent, how do we choose the *most likely* action sequences, when multiple are plausible? For example, for the task of selecting informational sources, how should we select sources for an article: will we prioritize a mix of "opposing" or "supporting" viewpoints (a stance-based plan)? Or a mix of "government", "academic" and "industry" voices? In [19, 8] (**Best Paper Award, C+J 2023**), I re-imagined classical methods for optimizing hidden variables to disambiguate plans: we determine that one plan is better than another if it better predicts the content we observe.

With the ability to detect high-quality, descriptive plans, we not only can derive fascinating insights into human workflows (see [13, 20, 12, 15]), but we can compare human plans with plans induced by prompting LLMs. In this vein, I made a significant finding: *AI plans do not resemble human plans and are less creative* [13]. This work, (**Outstanding Paper Award, EMNLP 2024**) is the first to show these shortcomings in creative settings and I have since replicated it in other fields [18, 5, 20]. *The fundamental insight is: current pretraining approaches do not lead to emergent human-level decision-making in AI*. **A major direction of my research going forward will be to build on these insights to train AI to behave in more human ways**. I plan to focus on developing and evaluating planning in real-world settings. I will compare offline evaluations of creative plans, conducted at scale [13], with online evaluations typical in Human-Computer Interaction (HCI) contexts. Additionally, I aim to broaden the scope of this research: my work has sparked a new subfield applying similar techniques across various domains, including law [18], creative writing [20], music, video script editing, and patent generation. Taken together, this body of work not only deepens our understanding of how planning drives creative processes but also opens the door to developing more adaptive and human-like AI systems.

During: LLM Generation with Adherence to Plans

In longer-form text generation, LLMs can easily drift away from their intended goals, especially when trying to incorporate detailed plans into the process. In this broad area of research, my focus is on ensuring that LLMs adhere more closely to predefined plans.

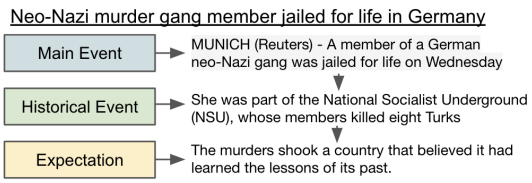

Fig. 2: We give an example of a latent plan (on the left) and an LLM adhering to that plan to generate output (on the right). This particular plan is a "discourse" plan, or a structural outline for how we wish to format the news article. We show in [11] that naive LLM output is not human-like: even the well-known Ovid's Unicorn article, despite being fluent, is structurally unnatural.

Once we have selected a plan, the next step is to integrate it into longer-form text generation. To address this, I introduced a concept I call structured generation, shown in Figure 2, which has since become a standard approach in fields (such as legal writing [24], instructional content [22], and story creation [6]). I tackled this challenge in two ways. First, in [11], I developed a method using a separate model, a trained classifier, to guide text. The classifier tells us the likelihood that a sentence belongs to the class of a certain structural element (e.g. "Main Event" or "Background", in Figure 2); I then use these likelihoods during generation to guide the sentence to look more like a certain class. Then, I introduced a novel post-generation classifier-guided editing step [11] to further enhance this pipeline. Running this approach sequentially, guiding each sentence towards a different structural element, as shown in Figure 2, we can induce a desired structure on the output.

Although this method requires additional training, it provides explainable and fine-grained control beyond other approaches. One interesting finding was that even low-accuracy classifiers could have a significant positive impact on controlling the text. This finding has implications for synthetic data creation and bootstrapped training, which I plan to explore further.

The second approach, done in collaboration with EleutherAI [16], involved adapting a technique called classifier-free guidance (CFG) and proving it could work in autoregressive contexts for text generation (**ICML 2024 Spotlight Award**). CFG queries a language model twice, generating two sets of next-token distributions: one made with the desired prompt, the other made without the prompt. Then, we subtract the unprompted distribution from the prompted distribution, making it *more likely* the next word will adhere to the prompt. This helps the model balance between adhering to a prompt and generating text naturally.

We found that, in addition to enforcing structural adherence, CFG in language modeling also reduces issues like hallucinations, improves accuracy, and helps introduce more structure into the generated text. This work has been adopted by OpenAI, integrated into major open source libraries like HuggingFace, and received an ICML 2024 spotlight award. There is much more to test with CFG: how to use CFG to adhere to a broader plan or structure; how well different parts of a prompt can be adhered to (e.g. in-context learning examples); how well we can how well we can "distill" CFG into a model by training further on CFG-generated synthetic data. I am currently working with a team at OpenAI to more fully explore this.

Looking forward, I would like to extend these methods to generate text with relation to more complex and real-world planning scenarios in order to work towards better and more faithful adherence to plans. My experiments thus far have only tested my methods in simplistic planning scenarios, like sentence-level story structure: what about more complex, hierarchal plans (i.e. document $\rightarrow$ paragraph $\rightarrow$ multi-sentence)? Or plans involving multiple dependencies or non-linear structures? I plan to integrate them into a more comprehensive and theoretically sound approach for both inferring plans and generating structured data. I believe that these methods can significantly enhance the effectiveness of language models.

After: Understanding AI's Societal Impact

While my work in the previous sections is aimed at delivering better tools for journalists and other creative professionals, there is always the risk that such work will be utilized by malicious actors to negatively impact society. This is where the third pillar of my work comes in. In [14], I worked with Eric Horvitz at Microsoft Research to understand the impact of Russian misinformation online. We utilized tweets, Facebook posts, Microsoft's web browser and search engine for a massive cross-platform study, to understand how users interacted with misinformation and the effects that it had. We were the first to show that misinformation was surfacing in search engine results, leading to considerable changes in Bing's design. However, such changes and continued vigilance is necessary. As techniques become more advanced and costs drop for generating more fluent, structurally sound content, such misinformation can become even more impactful and more challenging to detect.

When people are adversely affected by machine learning models, there needs to be a framework whereby individuals can quantify harm; here, defined as access to a resource (e.g. a "loan" or "probation"). In foundational work done in collaboration with Berk Ustin and Yan Liu [17], I developed the concept of actionable recourse, which measures both (1) whether a user can change a prediction made about them and (2) how much effort it takes. Some individuals might be denied a loan on the basis of an unchangeable attribute (e.g. gender or race) or a correlated attribute (e.g. income); in these circumstances, this individual will likely forever denied the loan. We developed a method to audit models based on what percentage of a population is denied recourse and to produce "flipsets", or actions that a user can take to reverse the prediction made about them. Our work – which has since been cited over 600 times, covered in Wired magazine and integrated into IBM AI Fairness 360 toolkit – has been foundational to this field.

In the future, I will continue to assess threats to our online ecosystem and take a cross-platform view that incorporates emerging modes of communication: new platforms (e.g. Telegram, Truth Social) and new mediums (e.g. podcasting). I also plan on continuing to develop methods for humans to seek redress for harms they have suffered. I want to extend the *recourse* framework to accommodate algorithmic exposure on social media platforms: I believe a broad connection. Put simply: when are humans exposed to different kinds of content, and can we use recourse methods to assess their recourse to altering this exposure pattern?

Future Research Agenda: More Faithful, Aligned AI with Safeguards Against Misuse

Looking forward, my dream is to integrate all three directions together to ultimately deliver **more aligned tools for creative endeavours** and **a cleaner information ecosystem**. I see the first two directions (i.e. Planning and Generation) coming together to help us develop and deploy creative tools that meaningfully improve human productivity. I hope to support journalists, especially local journalists, in producing more high-quality news at lower cost. Additionally, I will continue to expand this approach across creative domains, which is going to be a multi-year effort. Finally, the first two directions will improve the third (i.e. Misinformation/Recourse): work we have done in learning plans can help us better detect misinformation and misuse of LLMs. We have ongoing work looking at the sourcing patterns and article structure of *misinformation* compared with sourcing patterns in *mainstream news*. I believe that deeper structural analyses of text online is ultimately how we will continue combat misinformation, even as it continues to evolve, and assess harm (via recourse) of the humans who are exposed to it.

References

[1] **Alexander Spangher**, Kung-Hsiang (Steeve) Huang, Hyundong Justin Cho, and Jonathan May. Newsedits 2.0: Learning the intentions behind updating news. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2025.

[2] **Alexander Spangher**, Tenghao Huang, Yiqin Huang, Liheng Lai, Lucas Spangher, Sewon Min, and Mark Dredze. A novel multi-document retrieval benchmark grounded on journalist source-selection in newswriting. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2025.

[3] **Alexander Spangher**, Michael Lu, Hyundong Justin Cho, Weiyan Shi, and Jonathan May. Newsinterview: A dataset and a playground to evaluate llms' ground gap via informational interviews. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2025.

[4] **Alexander Spangher**, Nanyun Peng, Jonathan May, and Emilio Ferrara. Don't quote me on that: Finding mixtures of sources in news articles. In *Computation + Journalism*, 2020.

[5] Ryan Lee, **Alexander Spangher**, and Xuezhe Ma. Patentedits: Framing patent novelty as textual entailment. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2025.

[6] Dandan Li, Ziyu Guo, Qing Liu, Li Jin, Zequn Zhang, Kaiwen Wei, and Feng Li. Click: Integrating causal inference and commonsense knowledge incorporation for counterfactual story generation. *Electronics*, 12(19):4173, 2023.

[7] Alexander Spangher and Divya Choudhary. If it bleeds, it leads: A computational approach to covering crime in los angeles. *arXiv preprint arXiv:2206.07115*, 2022.

[8] Alexander Spangher, James Youn, Jonathan May, and Nanyun Peng. First steps towards a source recommendation engine: Investigating how sources are used in news articles. In *Computation + Journalism*, 2023.

[9] **Spangher, Alexander**, Emilio Ferrara, Ben Welsh, Nanyun Peng, Serdar Tumgoren, and Jonathan May. Tracking the newsworthiness of public documents. In *Proceedings of the 2024 Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

[10] **Spangher, Alexander**, Jonathan May, Sz-Rung Shiang, and Lingjia Deng. Multitask semi-supervised learning for class-imbalanced discourse classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 498–517, 2021.

[11] **Spangher, Alexander**, Yao Ming, Xinyu Hua, and Nanyun Peng. Sequentially controlled text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6848–6866, 2022.

[12] **Spangher, Alexander**, Nanyun Peng, Emilio Ferrara, and Jonathan May. Identifying informational sources in news articles. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

[13] **Spangher, Alexander**, Nanyun Peng, Sebastian Gehrmann, and Mark Dredze. Do llms plan like human writers? comparing journalistic coverage of press releases with llms. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

[14] **Spangher, Alexander**, Gireeja Ranade, Besmira Nushi, Adam Fourney, and Eric Horvitz. Characterizing search-engine traffic to internet research agency web properties. In *Proceedings of The Web Conference (WWW) 2020*, pages 2253–2263, 2020.

[15] **Spangher, Alexander**, Xiang Ren, Jonathan May, and Nanyun Peng. Newsedits: A news article revision dataset and a novel document-level reasoning challenge. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2022.

[16] **Spangher, Alexander**, Guillaume Sanchez, Honglu Fan, Elad Levi, and Stella Biderman. Stay on topic with classifier-free guidance. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

[17] **Spangher, Alexander**, Berk Ustun, and Yang Liu. Actionable recourse in linear classification. In *Proceedings of the 5th Workshop on Fairness, Accountability, and Transparency in Machine Learning (FAT/ML)*, 2018.

[18] **Spangher, Alexander**, Zihan Xue, Te-Lin Wu, Mark Hansen, and Jonathan May. Legaldiscourse: Interpreting when laws apply and to whom. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2024.

[19] **Spangher, Alexander**, James Youn, Matt DeButts, Nanyun Peng, and Jonathan May. Explaining mixtures of sources in news articles. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

[20] Yufei Tian, Tenghao Huang, Miri Liu, Derek Jiang, **Spangher, Alexander**, Muhao Chen, Jonathan May, and Nanyun Peng. Are large language models capable of generating human-level narratives? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

[21] Berk Ustun, **Spangher, Alexander**, and Yang Liu. Actionable recourse in linear classification. In *Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency (FAT)*, 2019.

[22] Te-Lin Wu, **Spangher, Alexander**, Pegah Alipoormolabashi, Marjorie Freedman, Ralph Weischedel, and Nanyun Peng. Understanding multimodal procedural knowledge by sequencing multimodal instructional manuals. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4525–4542, 2022.

[23] Te-Lin Wu, Caiqi Zhang, Qingyuan Hu, **Spangher, Alexander**, and Nanyun Peng. Learning action conditions from instructional manuals for instruction understanding. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023.

[24] Yang Zhong and Diane Litman. Strong–structure controllable legal opinion summary generation. *arXiv preprint arXiv:2309.17280*, 2023.