I seek to build more effective and human-aligned machine-in-the-loop systems by explicitly modeling human behaviors. I take cognitive science approaches, based on how humans learn from, or emulate [10], each other. As a former *New York Times* journalist, I saw firsthand how groups of journalists can come together to discover and disseminate high-quality information, positively impacting society. This process is informed by fuzzy, seldom-observed norms, and yet humans are able to perceive these and quickly agree on workflows together. **My research vision, inspired by these experiences, is to enable AI systems to do the same: to learn fuzzy human rewards, participate together in larger systems and become more effective collaborators.**

- **Modeling Human Rewards**: Current reasoning models under-perform for tasks with poorly-defined rewards [7] *(e.g. structuring narratives [24, 27, 29, 22] or combining multiple informational sources [21, 28])*. I develop methods to infer human goals from partial observations [18, 21, 31, 32, 28, 24], which I have used to train explicit planners [16]. As a faculty, I propose to extend this to (1) uncover human plans more efficiently using unsupervised latent variable estimation [8] (2) using these plans to infer human reward functions [1].
- **Plan-Aware Generation**: Even when provided with explicit directives, generative models may deviate from intended objectives. This poses a particular challenge when we incorporate more explicit plans into larger workflows. I have created techniques to *control* generative models and align them more closely to complex objectives [20, 19, 25]. I propose to extend this direction to develop agentic systems and that adhere better to human interventions.
- **Understanding AI's Societal Impact**: Once AI-generated content is disseminated, its effects 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 [23]. Additionally, I've examined how algorithmic decisions disproportionately affect certain demographics and proposed mitigation strategies [30]. Such methods are only increasing in importance the more we use algorithms to make decisions in our society. [30].

My work has been supported by **Stanford Artificial Intelligence Lab (SAIL)** and **Human-Centered AI (HAI)** Fellowships, a **4-year Bloomberg PhD Fellowship** and is being used at OpenAI, Google, 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 at EMNLP 2024 [22, 29], C+J 2023 [17], NAACL 2022 [24]), **Spotlight awards** (ICML, 2024 [25]), and **oral presentations** (NAACL 2024 [27]).

My work has been extended by peers in areas such as: scientific discovery, video script editing, creative writing, musical composition and legal writing. We have unified this field in a tutorial at NAACL 2025 called *"Creative Planning"* and are planning and proposing workshops for HCI, NLP and ML conferences. Finally, my work has gained substantial media attention, including by Wired and the New York Times.

Modeling Human Rewards

For models to be effective machine-in-the-loop partners in complex, creative tasks, they must be able to better understand, and then integrate, into human task flows. For instance, a journalist often follows multiple steps before drafting a news article (e.g. in Figure 1: "*call a source*" → "*call opposing source*" → "*find background*"). The core of my work here aims to model these task flows by developing novel techniques to *study human actions at scale*.

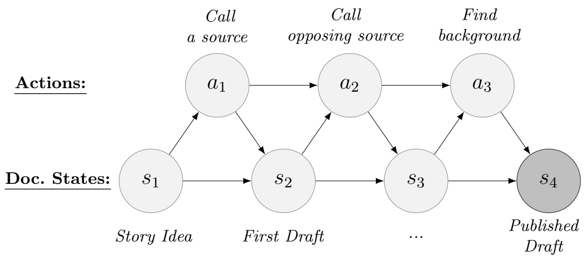


Fig. 1: Humans can infer the actions taken by humans (i.e. a_1, a_2..) simply by observing the final state of the process (s_4...), a phenomenon known to cognitive scientists as *emulation* [10]. Above we show a sample (state, action) sequence involved in generating a news article. This insight allows us to gather novel inferential data about human behaviors, infer human rewards and train agents to better integrate into creative processes.

Human workflows, as shown in Figure 1, can be modeled simply as (action, state) sequences. For many workflows *(e.g. in journalism [21], legal writing [27], etc.)*, data is often scare, and typically only the final state of this sequence is observable. This presents a particular challenge for training effective machine-in-the-loop systems. To overcome this challenge, we take inspiration from humans: *humans* have the ability to infer and learn from the actions of others simply by observing *this final state* – a process known to cognitive scientists as *emulation* [10]. For example, when we read a scientific paper, we, as researchers, infer details about implementation choices, negative trials, or hyperparameters, *even if unstated*.

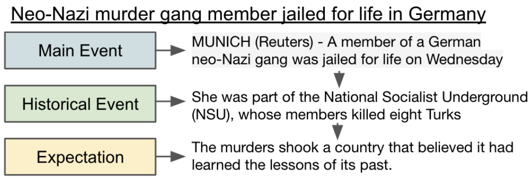
I have focused in my PhD on exploring this process: *using plentiful final-state data, I aim to (1) infer human actions, and (2)* *use these inferences to train more effective agents.* I have focused on tasks related to journalism, and have collaborated with professional journalists to study: how humans choose newsworthy stories [18], select sources [21, 28, 22, 4], and edit drafts [24, 3]. Methodologically, I designed Bayesian graphical models to infer *latent actions* from clusters of co-occurring words [6, 15]. I annotated datasets to train classifiers [18, 21, 27, 28, 19, 3]. In more recent work, I have developed novel approaches to use large language models (LLMs) to identify actions, yielding completely *unsupervised*, end-to-end action schemas [4, 5]. Finally, I showed that increasing visibility into intermediate states (i.e. drafts), can help us learn even more nuanced plans (**Outstanding Paper Award, NAACL 2022**) [24]. With the ability to detect high-quality, descriptive action sequences, we can compare human actions with plans induced by current models. In this vein, I made a significant finding: *current AI models do not resemble human planning processes* (Figure 1, [22]). 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 [27, 11, 29].

In summary, insights from my work thus far are: (1) human actions can be inferred simply from observing final documents and (2) current AI models fail to execute action sequences like humans. A major direction of my research going forward will be to build on these insights to train generative models to take more realistic human planning steps. I have three concrete directions I would like to take in my professorship.

Firstly, I would like to explore Bayesian Wake Sleep algorithm to infer action sequences [9]. A classical algorithm, Bayesian Wake Sleep aims to extend variational approaches for inferring latent variables by completely separating *model parameter inference* from *latent variable assignment*. I hypothesize that LLMs can provide an effective instantiations of this pipeline, allowing us to infer underlying structure of human actions in human data with greater accuracy. **Secondly**, I would like to use inferred human actions to distill human reward functions, using Inverse Reinforcement Learning (IRL) [2]. Although seldom applied to tasks in NLP, IRL gives us a powerful tool to train reward models and would allow us to depart from current practice of using large-scale preference-pair datasets to train reward models [14]. These reward functions, in turn, can help us train better systems, a process known as *apprenticeship* [1]. **Finally**, I aim to broaden the scope of this research: although my work has already extended beyond journalism *(e.g. in creative writing [29] and patent generation [12])*, I would like to apply these techniques more deeply in new domains.

Adhering to Larger Plans

Implementing longer-form plans, whether executed by a single LLM or by multiple collaborative agents, is still an area where language models struggle: generative models can easily drift away from their intended goals [20]. My focus here is on ensuring that AI adheres to user-specified plans.



MUNICH (Reuters) - A member of a German neo-Nazi gang was jailed for life on Wednesday

She was part of the National Socialist Underground (NSU), whose members killed eight Turks

The murders shook a country that believed it had learned the lessons of its past.

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 [20] that naive LLM output is not human-like: even the well-known Ovid's Unicorn article, despite being fluent, is structurally unnatural.

In my work, I introduced a concept called *structured generation*, shown in Figure 2, which has since become a standard approach in fields (such as legal writing [33], instructional content [31], and story creation [13]). I tackled this challenge in two ways. First, in [20], I developed a method using a separate planner to guide text, using the likelihood that the text belongs to a certain class to guide sampling probabilities. The second approach, done in collaboration with EleutherAI [25], involved adapting a technique called classifier-free guidance (CFG) and proving it could work in autoregressive contexts (**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 work has been adopted by OpenAI and integrated into major open source libraries like HuggingFace. 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.

Looking forward, I would like to extend these methods in several ways during my academic research going forward. *Firstly*, I wish to extend my approaches to orchestrate larger agentic workflows and enable 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 → paragraph → multi-sentence)? Or plans involving multiple dependencies or non-linear structures? **Secondly**, I would also like to explore how controlled generation can help us build better human-in-the-loop agentic flows. Currently, if a human takes an intervention in an agentic workflow, there is no principled way to adhere to that. However, with my tools we will be able to tune a system to be maximally responsive to the humans that use it. **Finally**, structured generative approaches give us a way to test *fairness* in a more principled manner. I will seek to build off of earlier work I did, developing *actionable recourse* [30]. Actionable recourse allows us to measure which features we need to change if we wish to change a prediction made by a classifier; this was foundational work I did in 2019 that led to the establishment of an entire field in FAccT research. I will seek to extend this work in the language modeling domain to test how a structured system can be altered to change it's output.

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 [23], 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 [26], 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 700 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 envision future work that develops more principled methods for connecting inferred human rewards with downstream system governance. Current approaches to AI alignment often rely on large-scale preference collection, but my research suggests that latent signals of reward can be extracted directly from the trajectories of professional practice. By coupling these signals with inverse reinforcement learning, I aim to produce reward models that more faithfully capture human intent and can serve as the basis for scalable oversight mechanisms. Such models could provide a foundation for aligning multi-agent systems where coordination, negotiation, and compromise are essential.

I plan to extend my agenda into the educational and participatory dimensions of AI research. Many of the domains I study—journalism, law, science—are also training grounds for future professionals. I hope to develop interactive platforms that allow students and practitioners to see, in real time, how their creative actions are modeled by AI systems, and to critique or refine those models. Embedding research insights into these collaborative environments will not only accelerate technical progress but also ensure that the systems I build are grounded in the lived experiences, values, and aspirations of the communities who will ultimately use them.

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] Pieter Abbeel and Andrew Y Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, page 1, 2004.

[2] Pieter Abbeel and Andrew Y Ng. Inverse reinforcement learning., 2010.

[3] **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.

[4] **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.

[5] **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.

[6] **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.

[7] Tuhin Chakrabarty, Philippe Laban, and Chien-Sheng Wu. Ai-slop to ai-polish? aligning language models through edit-based writing rewards and test-time computation. *arXiv preprint arXiv:2504.07532*, 2025.

[8] Kevin Ellis, Lionel Wong, Maxwell Nye, Mathias Sable-Meyer, Luc Cary, Lore Anaya Pozo, Luke Hewitt, Armando Solar-Lezama, and Joshua B Tenenbaum. Dreamcoder: growing generalizable, interpretable knowledge with wake–sleep bayesian program learning. *Philosophical Transactions of the Royal Society A*, 381(2251):20220050, 2023.

[9] Geoffrey E Hinton, Peter Dayan, Brendan J Frey, and Radford M Neal. The" wake-sleep" algorithm for unsupervised neural networks. *Science*, 268(5214):1158–1161, 1995.

[10] Lydia M. Hopper, Susan P. Lambeth, Steven J. Schapiro, and Andrew Whiten. Observational learning in chimpanzees and children studied through 'ghost' conditions. *Proceedings of the Royal Society B: Biological Sciences*, 275(1636):835–840, 2008.

[11] 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.

[12] Ryan Lee, Alexander Spangher, and Xuezhe Ma. Patentedits: Framing patent novelty as textual entailment. *arXiv preprint arXiv:2411.13477*, 2024.

[13] 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.

[14] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[15] 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.

[16] Alexander Spangher, Tenghao Huang, Yiqin Huang, Lucas Spangher, Sewon Min, and Mark Dredze. A novel multi-document retrieval benchmark: Journalist source-selection in newswriting. In *Proceedings of the Fourth Workshop on Knowledge Augmented Methods for NLP (KAMNLP), Nations of the Americas Chapter of the ACL*, 2025.

[17] 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.

[18] **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.

[19] **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.

[20] **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.

[21] **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.

[22] **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.

[23] **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.

[24] **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.

[25] **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.

[26] **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.

[27] **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.

[28] **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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

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