Thesis Proposal: Detecting Agency Attribution

Igor Ryazanov and **Johanna Björklund**
Umeå University, Department of Computing Science
[igorr, johanna]@cs.umu.se

Abstract

We explore computational methods for perceived agency attribution in natural language data. We consider 'agency' as the freedom and capacity to act, and the corresponding Natural Language Processing (NLP) task involves automatically detecting attributions of agency to entities in text. Our theoretical framework draws on semantic frame analysis, role labelling and related techniques. In initial experiments, we focus on the perceived agency of AI systems. To achieve this, we analyse a dataset of English-language news coverage of AI-related topics, published within one year surrounding the release of the Large Language Model-based service ChatGPT, a milestone in the general public's awareness of AI. Building on this, we propose a schema to annotate a dataset for agency attribution and formulate additional research questions to answer by applying NLP models.

1 Introduction

The value of studying power relations through the lens of language has been investigated in various contexts from online communities to film plots, see e.g. Bramsen et al. (2011); Danescu-Niculescu-Mizil et al. (2013); Sap et al. (2017). Across different fields, agency has a range of definitions, which highlight different aspects of the concept. Within the context of this work, agency is taken as the freedom and capacity of an entity to act, corresponding to one facet of power. *Perceived agency* is then the agency that we project on other entities while interpreting a description of a situation. Perceived agency is important because it signals autonomy and independence, but also moral accountability: it is hard to imagine a hero or villain who is always a victim of circumstance. The perception of agency also influences how we assign blame or praise. For example, the *actor-observer* cognitive bias (Jones and Nisbett, 1971) is the general tendency to explain other individuals' behaviours as an effect of their personalities, i.e. as something they cannot help doing given who they are, and our own behaviours as the rational response to our current situation. For example, someone else's slow driving may be attributed to their age or gender, but when we drive slowly we attribute it to specific reasons such as worn tyres. If we feel we deserve credit, we can frame our behaviour as an active choice; if we want to avoid guilt, we can emphasise the external pressure.

To assign, e.g. responsibility, agency attribution is frequently manipulated in political discourse and partisan reporting to affect the audience (Iyengar, 1994). For example, there are discursive techniques to humanise or dehumanise migrants that draw on agency (Kirkwood, 2017): portraying incoming migrants as independent agents (e.g. asylum seekers) in opposition to more passive roles (refugees) naturally affects public perception and can influence the assumptions underlying political decisions (Sajjad, 2018). However, even without an explicit political intention, the wording may suggest how agency is assigned. Wikipedia, for example, is widely known to have its guidelines built around a 'neutral point of view', but the editorial bias (not limited to agency) that the guidelines seek to eliminate remains (Hube, 2017). In the general audience media, it is not uncommon to observe agency shifts between collectives and individuals ('The company has decided to lay off' vs. 'The board has decided' vs. 'The CEO has decided') and between individuals and artefacts ('The car crashed into the bridge' vs. 'The driver crashed his car into the bridge') (Te Brömmelstroet, 2020). Furthermore, perceived agency plays a key role when discussing entities that are specifically designed to appear intelligent. Most prominently – the various technologies referred to as AI, but also, for example, toys, voice recognition systems, and non-playable video game characters. Therefore, perceived agency is

Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop, pages 208–214
March 21-22, 2024

created by a mixture of intentional and unintentional messaging.

In this proposal, we focus on perceived agency as a prediction target for computational models, and as a direction for linguistic analysis in computational social sciences. If framed as a question-answering task, the prompt would be 'Who or what acts or can act independently and intentionally in this situation?'. From this, we derive two research objectives. The first is to investigate the computational approach to studying attributed agency (RQ1), in particular, the efficacy of different computational models in predicting perceived agency.

RQ1: What linguistic features and computational models are most suitable for predicting perceived agency?

The second goal is to apply NLP methods as part of studies in computational social science to understand agency in specific contexts. We are interested in learning whether: (i) it is feasible to use such computational models to measure the public's perception of agency, (ii) the models generalise across narrow topics, and (iii) these models can be part of topic-specific social scientific studies and combined with qualitative approaches. In short:

RQ2: To what extent can automated prediction of perceived agency in text answer questions from social science? What questions can it answer?

2 Prior computational work

There is limited research on NLP methods for agency attribution. The closest work is likely that by Minnema et al. (2022a) on *perceived responsibility*, which is part of their broader work on detecting perspectivisation (Minnema et al., 2022b) using FrameNet-based annotations (Baker et al., 1998). They project perceived responsibility in a sentence onto three axes: blame, focus, and cause, and demonstrate that text features account for some of the differences in these facets of perceived responsibility. In particular, they study Italian news reporting of femicides and Dutch articles reporting traffic accidents (Minnema et al., 2022b,a). The latter is based on the study by Te Brömmelstroet (2020) that investigates news headlines about traffic accidents. The authors annotated articles into agentive and non-agentive categories based on the phrasing. This group of works focus on the attribution of *responsibility* ('Who or what is blamed in this situation?'), rather than specifically on agency as we describe it above.

A larger group of related work is focused on the (perceived) semantic roles that entities play in a text. This approach goes back to the folklorist studies of Propp (1968) and focuses on identifying archetypes, such as 'hero', 'villain' or 'victim'. Computational applications of Propp's ideas vary from directly applying components of his grammar to new texts (Finlayson, 2016) to using Large Language Models (LLMs) for zero-shot role labelling (Stammbach et al., 2022). In terms of domains of application, news articles are prominent (Stammbach et al., 2022; Gomez-Zara et al., 2018), but role prediction has also been applied in other settings containing political discussions. For example, identifying semantic roles in memes has been used as a shared task and prompted both text-based and multimodal solutions (Sharma et al., 2022). Finally, semantic role labelling, as well as FrameNet-style annotations, are used in the field of emotion detection (Bostan et al., 2020). As in agency attribution, emotion detection assumes a choice of perspective (i.e. that of the writer or reader) before making predictions.

A particularly relevant study by Sap et al. (2017) is related to both of these groups of works and introduces frames of agency and power to investigate the subtler types of gender bias in modern films. Their study focuses on establishing the agency of characters throughout a longer narrative (compared to shorter messages we are interested in) and emphasises authority as one of the main indicators of agency, but even with these differences in approach, it remains one of the closest points of reference for this proposal.

3 Application areas

We consider two application areas, namely the agency ascribed to AI systems and the examination of bias in news reporting.

AI anthropomorphism Our first domain of interest, to which most of our preliminary work has been dedicated, is the ongoing discourse on systems claiming to be artificial intelligence (AI). Recent developments of LLMs and their branding as 'AI' reinforce the anthropomorphisation of the technology. Generative models, especially those used in chatbots, tend to emulate first-person human speech, and end users are intended to project higher levels of agency on these systems. This may have positive effects in some contexts (Sheehan et al., 2020) but also have highly adversarial effects, e.g. a dangerous over-reliance on the system (Abercrombie et al., 2023).

More broadly, the degree to which AI systems are viewed as active agents is reflected in public conversation and news coverage. The impact of alternative phrasings is illustrated in Table 1. In the first example, agency shifts in a similar way to what has been observed in the news coverage of car crashes (Te Brömmelstroet, 2020), in the sense that an instance of a technological artefact is used to refer to people or organisations (AI companies, in this case). It stands, however, to reason that, unlike in the case of cars, the word choice is not purely rhetorical because people may perceive true agency from 'AI' actors (an observation which deserves further investigation). The second example in Table 1 provides one depiction of an AI system as a conversational partner and one as a tool. The agency is thus ascribed either, at least in part, to the system or to the self in full.

Reporting bias Our second domain of interest is attributed agency bias towards marginalised groups in news reporting. Rhetoric plays a significant role in advancing political agendas and through the "correct" linguistic choices, stereotypical qualities can be ascribed to individuals. For example, the labelling of individuals as opportunistic or immoral has been demonstrated to influence public opinions and migration policies (Kirkwood, 2017; Findor et al., 2021; Sajjad, 2018). Agency attribution is an important aspect of this phenomenon and a focus of our investigation.

In both of these domains, language reflects how the media or the public interprets specific technological and social issues. These interpretations by the mass media inform and influence those who make policies and regulations, thus translating perceptions into reality.

4 Proposed work

This section outlines the preliminary and proposed future work that will go into the thesis.

4.1 Preliminary work

In an ongoing study of AI in news reporting, we analyse the descriptions of AI systems performing various tasks. Since the release of the LLM-based ChatGPT in November 2022, there has been a massive increase in publications on AI. Interestingly, the statistics reported by Google Trends (Figure 1) indicate that previous releases of generative tools such as Stable Diffusion, DALL-E 2, and Midjourney (all made in the summer of 2022) did not correspond to any significant increase in the general public's interest in AI. The quantitative change in AI news coverage was only brought on when LLMs entered the scene. Our research goal for this study is to investigate whether the use of the term 'AI' has changed qualitatively as well. In other words, whether journalists write about AI differently now compared to the time before ChatGPT.

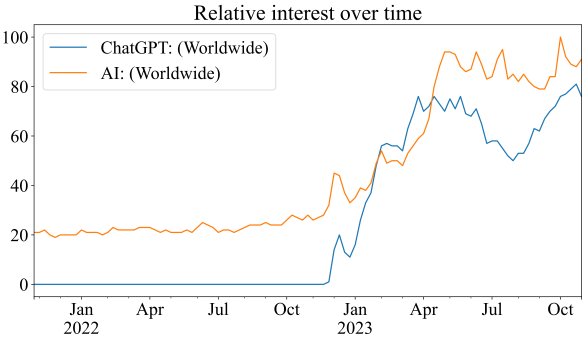

Figure 1: Relative interest in 'ChatGPT' and 'AI' between October 2021 and October 2023 based on search queries (normalised search numbers over the period). Source: Google Trends (https://trends.google.com/trends/).

For the study, we collected a dataset of 6 150 articles mentioning the term 'AI'. We choose to work with general-domain publications because we expect these to better reflect the language of the general public than, for example, technical or scientific publications. For legal reasons, the set of publications was restricted to such that do not impose paywalls. The dataset covers publications from May 31, 2022, to May 31, 2023, and contains 19 out of the 25 largest English-language news websites as of May 2023 (Majid, 2023). As part of our analysis, we assign FrameNet annotations to sentences mentioning 'AI' using the FrameNet parser from the information extraction system LOME (Xia et al., 2021). The annotated frames provide a convenient way to study the uses of 'AI' by the type of situation at a relatively large scale. For the purpose

High communicated agency	Low communicated agency
‘Will AI steal your job?’	‘Will the AI companies disrupt the job market?’
‘AI helped me with my homework.’	‘I used AI tools in my homework.’

Table 1: The impact of phrasing on the communicated agency of AI systems.

of this proposal, we are interested in frames and their attributes that may indicate that some level of agency is assigned to ‘AI’.

A number of FrameNet frames that describe cognitive efforts implying at least some degree of agency (e.g. `awareness`, `coming_to_believe`, `opinion`) have an attribute `cognizer` that refers to a sentient being enacting these efforts. Selecting such frames where AI plays the role of a ‘cognizer’ provides us with a subset of data where it is likely to have some perceived agency. Out of 609 such occurrences in our dataset, the most common constructions involve AI ‘thinking’ (87 instances), ‘analysing’ (47), ‘making decisions’ (39), ‘predicting’ (37) and ‘learning’ (30). Thinking is most often, but not always, brought up in articles such as: ‘Here’s how AI thinks X would look’ (Table 2), clearly anthropomorphising an AI system and giving it intentionality. In the same subset of frames, other constructions are often used to describe the normal functionality of a system relatively neutrally, e.g. ‘...AI is good at recognizing patterns...’ (Table 2). In this case, the perception of agency behind AI would arguably be lower than in the one above. These examples illustrate how, even within semantically similar constructions, implied and perceived agency can differ significantly. They also demonstrate that while evoking specific frames does not necessarily correspond to agency directly, pipelined FrameNet annotations have their use in identifying descriptions of situations with ambiguous agency.

4.2 Phase 1: Dataset annotation

Our first goal is to define the computational task for perceived agency detection. For this, we need an annotated dataset that covers two or more domains to ensure we can understand how well the solutions generalise. The first area of interest was discussed in Section 4.1: the perceived agency of AI in mass media coverage. We plan to annotate various statements from AI and technology-related articles published by mainstream media derived from the news corpus described in the same section. One of the goals of our study is to compare perceived human and technological agency, and the second part of the new dataset will be focused on the portrayal of humans in newspaper headlines, with a balance of topics such as politics, entertainment, crime, etc.

Through the crowdsourced annotation process, we aim to both create a dataset fitting for the perceived agency detection and investigate how the annotation reflects annotators’ interpretation of the topic. We interpret agency as the capability to take intentional actions and, even more broadly, influence the situation. In the experimental setting, a ‘situation’ is fully described in one or several sentences (e.g. by a news headline or a paragraph) and should contain only several entities displaying agency. Therefore, we consider it a reasonable annotating task to rate the degree of agency exhibited by these entities according to a reader’s perception.

We propose the following annotating process (examples of steps 1 and 2 in Table 3):

1. **Identification.** Annotators are given a broad explanation of our interpretation of agency and asked to highlight all entities that have agency in the described situation, with an option to write in external entities.
2. **Specification.** For each entity, the respondents answer a multiple-choice question about the level of agency the entity has in the situation (‘How would you describe the agency the X demonstrates in this situation?’). The degrees of control given in the answers are ‘complete control’, ‘a high level of control’, ‘some or shared control’, ‘little control’, and ‘no control at all’.
3. **Resolution.** To resolve annotation conflicts, we propose using the longest, most common, subsequence rule (Bostan et al., 2020) for highlighted entities.
4. **Aggregation** To aggregate the multiple-choice answers, the annotations can be converted from categories to numerical values. This can be done by assigning numerical values to the possible answers and computing average scores for annotated entities. Because the scores represent the

Table 2: Examples of sentences with LOME-labelled frames where AI plays the role of a ‘cognizer’.

Remember, AI is good at recognizing patterns, and humans are good at understanding when those patterns have meaning versus when they are spurious correlations.	Here’s what AI thinks Barbie will look like at ages 50, 60, and 70.
Frame: ‘Becoming_aware’ – recognizing Attributes: ‘Cognizer’ – AI ‘Phenomenon’ – patterns	‘Awareness’ – thinks ‘Cognizer’ – AI ‘Content’ – what ‘Content’ – Barbie will look like at 50, 60, and 70.

Table 3: Proposed annotation example. As a first step, annotators are offered to highlight entities with agency. The second step is annotating the level of agency for each entity: from low (recognised as an agent but next to no influence) to complete (full control of the situation).

Sentence	Mary asked AI for help with her homework.	Mary used an AI tool in her homework.
First step Entities with agency	**Mary** asked **AI** for help with her homework.	**Mary** used an AI tool in her homework.
Second step Level of agency	**Mary** asked **AI** for help with homework **Mary** – high, **AI** – medium	**Mary** used an AI tool in her homework. **Mary** - complete

agency in a specific closed situation, they should be normalised over the situation.

Step 1 can be complemented by named-entity recognition to identify noun phrases not marked by the annotators, resulting in a category for the entities that are not even considered to be agents, i.e. incapable of taking an active role at all (as opposed to the ones perceived as agents but considered not to have agency in the specific situation).

4.3 Phase 2: Predicting perceived agency level

Unlike stance or opinion mining, which require broad semantic context, the level of agency can be expressed with shorter spans of text and more syntactical instruments, such as passive voice or the choice of a specific synonym. Therefore, it is reasonable to expect machine learning approaches to perform well on agency attribution, even when applied to single sentences. This assumption is supported by Minnema et al. (2022a), who reported encouraging performance of a fine-tuned BERT-based model on a similar, but even more topic-specific and granular, perception mining task. Based on this, we are aiming to test several models of different levels of complexity on the annotated dataset. In particular, through our experiments, we are interested in answering the following questions:

- Can a pre-trained language model (e.g. BERT*) be fine-tuned to predict perceived agency? If yes, would fine-tuning such a model on a dataset covering one topic (e.g. AI news) transfer the performance to another (e.g. culture news)?
- Can existing named entity recognition models or more generalised semantic information extraction models, such as LOME, be directly useful in predicting perceived agency?
- Is it possible to reliably annotate perceived agency with LLMs so that the result is consistent with human judgement?

5 Conclusion

When we interpret stories, make decisions based on them, or place responsibility or blame, we rely on our perception of agency to understand whose intentions are driving events. Language choices can intentionally and unintentionally influence this perception and, ultimately, our reactions. In this proposal, we put forward perceived agency detection as an NLP task and outline our preliminary and planned work on creating and annotating a perceived agency dataset. Our focus is on two topical areas: a narrow one (perceived agency of AI) and a broader one (perceived agency in news headlines, with an eye towards bias). We describe some of our planned computational experiments, which will evolve as we learn from our findings, and aim for computational social science applications. We hope this proposal brings focus to the notion of perceived agency and highly welcome all types of feedback to further improve it.

Limitations

As with any study based on collecting data on human perceptions from a limited number of participants, the proposed thesis relies heavily on the assumption that the surveyed demographics and their responses are sufficiently representative to make results generalisable. Both the dataset and the choice of annotator will inevitably introduce bias that needs to be considered and reported. Due to how the annotated data is acquired, we further limit our definition of agency. For example, besides studying the perceived agency of entities mentioned in a message or the messenger, it is worth considering the perceived agency of the readers themselves, which we do not investigate here. In particular, in contexts with frequent direct messaging, such as advertising and political communication, influencing the readers' sense of agency can be a nudging or manipulating technique. However, assessing participants' self-perception based on texts would likely require a different set of tools, as well as considerable expertise in psychology. Similarly, it can be argued that the source of information (e.g. a news article vs. a social media post vs. a generated response by a chatbot) may affect how humans perceive it. However, within the framework of this project, we do not yet have the means of assessing the influence of the text source on human interpretation of agency.

Another significant limitation is that the current proposal is limited only to English-language media, largely due to their international dominance. If the perceived agency prediction task is reliably solvable, it should be further considered in the multilingual setting. As shown by, e.g. Findor et al. (2021), types of agency and perceptions can shift significantly through literal translation because of different etymologies and connotations. Therefore, building a multilingual corpus out of direct translations that imply different levels of agency may present a more challenging task.

Acknowledgements

This research is supported by the Wallenberg AI, Autonomous Systems and Software Program – Humanities and Society (WASP-HS) and the Marianne and Marcus Wallenberg Foundation. We thank Hannah Devinney, who shared their thoughts at the very start. We also thank the reviewers of the 2024 EACL Student Research Workshop, whose insightful comments helped improve this proposal as well as its implementation.

References

Gavin Abercrombie, Amanda Curry, Tanvi Dinkar, Verena Rieser, and Zeerak Talat. 2023. Mirages. on anthropomorphism in dialogue systems. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4776–4790, Singapore. Association for Computational Linguistics.

Collin F Baker, Charles J Fillmore, and John B Lowe. 1998. The Berkeley FrameNet project. In *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*.

Laura Ana Maria Bostan, Evgeny Kim, and Roman Klinger. 2020. GoodNewsEveryone: A corpus of news headlines annotated with emotions, semantic roles, and reader perception. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1554–1566, Marseille, France. European Language Resources Association.

Philip Bramsen, Martha Escobar-Molano, Ami Patel, and Rafael Alonso. 2011. Extracting social power relationships from natural language. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 773–782, Portland, Oregon, USA. Association for Computational Linguistics.

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A computational approach to politeness with application to social factors. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 250–259, Sofia, Bulgaria. Association for Computational Linguistics.

Andrej Findor, Matej Hruška, Petra Jankovská, and Michaela Pobudová. 2021. Re-examining public opinion preferences for migrant categorizations: "refugees" are evaluated more negatively than "migrants" and "foreigners" related to participants' direct, extended, and mass-mediated intergroup contact experiences. *International Journal of Intercultural Relations*, 80:262–273.

Mark Alan Finlayson. 2016. Inferring Propp's functions from semantically annotated text. *Journal of American Folklore*, 129(511):55–77.

Diego Gomez-Zara, Miriam Boon, and Larry Birnbaum. 2018. Who is the Hero, the Villain, and the Victim? Detection of roles in news articles using natural language techniques. In *23rd International Conference on Intelligent User Interfaces*, IUI '18, pages 311–315, New York, NY, USA. Association for Computing Machinery.

Christoph Hube. 2017. Bias in Wikipedia. In *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW '17 Companion, page 717–721, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Shanto Iyengar. 1994. *Is anyone responsible?: How television frames political issues*. University of Chicago Press.

Edward E Jones and Richard E Nisbett. 1971. The actor and the observer: Divergent perceptions of the causes of behavior. In *Attribution: Perceiving the causes of behavior*.

Steve Kirkwood. 2017. The humanisation of refugees: A discourse analysis of UK parliamentary debates on the European refugee 'crisis'. *Journal of Community & Applied Social Psychology*, 27(2):115–125.

Aisha Majid. 2023. Top 50 biggest news websites in the world. Press Gazette, [Accessed: 12/12/2023].

Gosse Minnema, Sara Gemelli, Chiara Zanchi, Tommaso Caselli, and Malvina Nissim. 2022a. Dead or murdered? Predicting responsibility perception in femicide news reports. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1078–1090, Online only. Association for Computational Linguistics.

Gosse Minnema, Sara Gemelli, Chiara Zanchi, Tommaso Caselli, and Malvina Nissim. 2022b. SocioFillmore: A tool for discovering perspectives. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 240–250, Dublin, Ireland. Association for Computational Linguistics.

Vladimir Propp. 1968. *Morphology of the Folktale*. University of Texas Press.

Tazreena Sajjad. 2018. What's in a name? 'Refugees', 'migrants' and the politics of labelling. *Race & Class*, 60(2):40–62.

Maarten Sap, Marcella Cindy Prasettio, Ari Holtzman, Hannah Rashkin, and Yejin Choi. 2017. Connotation frames of power and agency in modern films. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2329–2334, Copenhagen, Denmark. Association for Computational Linguistics.

Shivam Sharma, Tharun Suresh, Atharva Kulkarni, Himanshi Mathur, Preslav Nakov, Md. Shad Akhtar, and Tanmoy Chakraborty. 2022. Findings of the CONSTRAINT 2022 shared task on detecting the hero, the villain, and the victim in memes. In *Proceedings of the Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situations*, pages 1–11, Dublin, Ireland. Association for Computational Linguistics.

Ben Sheehan, Hyun Seung Jin, and Udo Gottlieb. 2020. Customer service chatbots: Anthropomorphism and adoption. *Journal of Business Research*, 115:14–24.

Dominik Stammbach, Maria Antoniak, and Elliott Ash. 2022. Heroes, villains, and victims, and GPT-3: Automated extraction of character roles without training data. In *Proceedings of the 4th Workshop of Narrative Understanding (WNU2022)*, pages 47–56, Seattle, United States. Association for Computational Linguistics.

Marco Te Brömmelstroet. 2020. Framing systemic traffic violence: Media coverage of Dutch traffic crashes. *Transportation Research Interdisciplinary Perspectives*, 5:100109.

Patrick Xia, Guanghui Qin, Siddharth Vashishtha, Yunmo Chen, Tongfei Chen, Chandler May, Craig Harman, Kyle Rawlins, Aaron Steven White, and Benjamin Van Durme. 2021. LOME: Large ontology multilingual extraction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 149–159, Online. Association for Computational Linguistics.