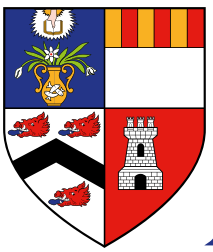
Cognitive Rigidity as a Moderator of LLM-Social Influence in Political Contexts

Alexander Probst

Dr Kevin Allan

3rd of August, 2025

Words: 7259 (Intro+Discussion)



UNIVERSITY OF ABERDEEN

Cognitive Rigidity as a Moderator of LLM-Social Influence in Political Contexts

Table of Contents

Abstract

Large Language Models (LLMs) are a rapidly evolving technology. They are getting increasingly integrated into people's lives, including democratic processes. At the same time, the understanding of the underlying mechanisms of LLM-persuasion compared to human persuasion, and scientific means to examine it, are still limited. The aim of this study was to investigate the potential of using psychometric scales to define LLM agents as experimental stimuli, and to use said agents to examine the role of cognitive rigidity in human-ai influence dynamics. In a human-ai interaction study, liberal and conservative LLM agents, created through what we call "psychometric prompting", had the goal to persuade participants with varying degrees of cognitive rigidity of their political viewpoints on social and economic issues. We hypothesised that participants would generally be influenced by agents, that they would be influenced more by politically aligned than by misaligned agents, and that cognitive rigidity would moderate said influence. We found initial support for psychometric prompting, but not for the expected effects of alignment and cognitive rigidity. We conclude that while the current results are incongruent with contemporary literature, further research addressing limitations of the current study is needed to arrive at a conclusive contribution on the persuasiveness of LLM agents and the role of cognitive rigidity in human-ai influence dynamics.

Introduction

Over the last few years, concerns have grown about the persuasive power of large language models (LLMs) (Bengio et al., 2024; Burtell & Woodside, 2023; El-Sayed et al., 2024; Floridi, 2024; Grace et al., 2024). This concern mainly stems from mounting demonstrations of the persuasive capabilities of LLMs (Allan et al., 2024; Breum et al., 2024; Dillion et al., 2025; Hackenburg et al., 2023a; Hackenburg & Margetts, 2024; G. Huang & Wang, 2023; Salvi et al., 2024; Schoenegger et al., 2025; Shi et al., 2020; Song et al., 2024) across many domains (including politics). It also stems from growing evidence of unintentional bias in the models through their training data (Crawford, 2021; Lin et al., 2024; Oketunji et al., 2023; Orr & Crawford, 2024; Zook et al., 2017), and the increasing ease with which models can be given intentional biases and be deployed for interaction with people at scale (Cirulli et al., 2025; Hackenburg et al., 2023a). Particularly in domains like politics, the possibility of systematic and

intentional bias, exhibited by LLMs actively used to persuade people, seems particularly alarming. This concern gets contextualised by the seeming normalization of this novel technology within democratic processes. One example is the debut of government and non-government LLM-driven applications, designed to help people decide who to vote for by chatting about parties and their programmes, in the German 2025 general election(Hesselmann, 2025; Schiffer, 2025).

While research in this area grows quickly, there are still many gaps. Human dispositional factors (e.g. personality traits, epistemic and affiliative needs) are known to play important roles in human-human persuasion (Jost, 2009; Jost et al., 2009; Petty & Cacioppo, 1986). In the human-ai context, however, they are currently underexamined. Additionally, Human-ai interaction research, including that on political persuasion, increasingly uses LLM agents as experimental stimuli (Allan et al., 2024; Costello et al., 2024; Hackenburg & Margetts, 2024; Ziems et al., 2024). Agents are created through prompting, which involves giving instructions to the model as to how to behave when talking to a user. One challenge with this in current human-ai interaction research is coarse behavioural control of agents (Sclar et al., 2024; Wang et al., 2023). While lots of different prompting strategies are applied in various ways, the search for a shared framework for maximising control, consistency and replicability, is still ongoing. These two themes, understanding the influence dynamic between humans and LLM agents better, particularly the role of human dispositional factors, and contributing towards an established framework using LLM agents as experimental stimuli, is what we are trying to address with this study.

1.1 Beliefs and Attitudes

Beliefs and attitudes are foundational for most behaviour, and certainly for political action such as voting or protesting. Beliefs stand for a person's thoughts associated with a specific object of knowledge (Huddy et al., 2023). They can be of an evaluating (I believe that abortion clinics are good because they represent women's self-determination) or non-evaluating (I believe that there are 55 reproductive health care clinics in the UK) nature. Attitudes are evaluative summaries associated to objects of knowledge (e.g., I feel positive about abortion clinics) (Huddy et al., 2023). They can be positive or negative, strong or weak, as well as implicit and explicit. Beliefs form the basis for ideologies. While there is historical incongruence about their definition, an ideology is, in the most broad sense, simply a set of beliefs about any given topic. (Jost et al., 2009). A political ideology could, therefore, for example be a set of beliefs about society and how it should operate.

In a given individual, beliefs and attitudes are not static – people at voting age aren't blank slates – but are rather dynamic and evolving products of constant information processing happening in everyday life (Evans & Stanovich, 2013; Huddy et al., 2023; Kruglanski & Gigerenzer, 2011; Petty & Cacioppo, 1986). Research suggests that the kind of political beliefs, attitudes and ideologies people adopt, aren't arbitrary. Jost et al (2003; 2009) argue that dispositional traits, like death anxiety, a need for cognition (the tendency to enjoy thinking extensively about something) and tolerance of ambiguity, are predictive of whether people tend to adopt conservative or progressive beliefs, attitudes and ideologies. Such dispositional factors (including but not limited to the ones just mentioned) also play an important role in theories of information processing. This is because, together with contextual factors (Huddy et al., 2023), they influence an individual's motivation to engage with incoming information, for example whether they spend more or less cognitive effort on processing (peripheral vs elaborative) or whether that processing is directionally motivated (e.g. refuting incongruent information through cognitively expensive counterarguing) (Evans & Stanovich, 2013; Kruglanski & Gigerenzer, 2011; Petty & Cacioppo, 1986). How individuals engage with incoming information in turn predicts what aspects of a message (e.g. the status of the messenger, or the quality of the argument) can lead to a change in beliefs and attitudes and whether the resulting beliefs and attitudes are more fickle or stable. More stable beliefs are a) more predictive of behaviour and b) harder to change (Huddy et al., 2023). Another relevant aspect in the political context is that of social identity. Social identity theory applied to politics describes how a social group we associate with, e.g. a political party, and the strength of that association, e.g. a partisan, can in any given situation alter our priorities and motivations if it is activated (Ellemers et al., 2001; Huddy, 2001; Tajfel & Turner, 2004). Political beliefs, attitudes and how they are updated, is tied to the political group people associate with. A political identity constitutes another motivational factor, influencing processing by for example causing ego and group-justifying reasoning (e.g. when confronted with people or messages representing one's political out-group) (Jost et al., 2022; Petty & Cacioppo, 1986).

In other words, people have a tendency, based on their individual expression of fundamental dispositional traits, to 'naturally' like certain (political) ideas better than others. These traits also, once there are existing beliefs and an identity around them, shape (alongside many other contextual factors) how people engage with new information that is congruent or incongruent with their existing beliefs (Huddy et al., 2023; Petty & Cacioppo, 1986). Messages from a LLM on political topics would be an example of such incoming information. While it is

important to note that the mentioned contextual factors play an important role too, within the scope of the current study we focus on the motivational role of dispositional factors, as this is what we try to experimentally isolate and examine.

1.1.1 A Variable of Special Interest – Cognitive Rigidity

One example of just mentioned dispositional factors is called cognitive rigidity (CR). CR is defined as a person's domain general (in)ability to adapt thinking patterns and problem solving strategies in the face of novel information, particularly when it presents unwise to stick to previous strategies and patterns in light of said novel information (e.g., when novel information changes the factual situation of a matter which in turn requires re-evaluation to maintain accurate judgement) (Zmigrod, 2020). Cognitive rigidity can be measured through self-report and cognitive behavioural tasks (Jost, 2021; Van Hiel et al., 2016; Zmigrod et al., 2019). One common self-report measure is the cognitive flexibility inventory (CFI) (Dennis & Vander Wal, 2010). Common cognitive behavioural measures include the Wisconsin Card-sorting task (WCT) (Grant & Berg, 1948) or task switching paradigms (Miyake & Friedman, 2012; Monsell, 2003). Notably, the kind of measure one applies seems to be having a particular effect. When measuring cognitive rigidity through self-report, evidence suggests that it is more prevalent among conservatives (Jost, 2021; Jost et al., 2003, 2007). When measured through behavioural tasks, it seems like rigidity is prevalent among extremists from both sides of the political spectrum (Greenberg & Jonas, 2003; Van Hiel et al., 2010, 2016; Zmigrod et al., 2019)

CR is a trait of particular interest in the context of political beliefs and attitudes. This is because it reflects key dispositional factors that influence, as described above, how a person engages with incoming information that is aligned or misaligned with prior beliefs. For instance, CR is positively correlated with increased motivated reasoning behaviours, such as confirmation and disconfirmation bias and selective exposure (Jost et al., 2022; Krems, 2014). Further, it is linked to ideological extremism, particularly via its positive correlation with dogmatism and authoritarianism (both are often used as trait-measures of CR) (Fransen et al., 2015; Zmigrod, 2020; Zmigrod et al., 2020). In other words, people high in cognitive rigidity tend to *disproportionately* a)expose themselves to information that confirms their prior beliefs while avoiding the opposite and b) treat information confirming their views more favourably than contradicting information, *compared to* people who are more cognitively flexible (Huddy et al., 2023; Jost et al., 2022; Petty & Cacioppo, 1986). An example would be a conservative, who is anti-welfare benefits, accepting superficial arguments against increased social spending from

a politician of their party at face value, while expending lots of cognitive resources to counterargue more sophisticated suggestions in favour of social spending by a liberal politician. A practical consequence of this can be radicalization. A cognitively rigid person is, considering the evidence, arguably more likely to radicalize, for instance in an online environment where they are disproportionally exposed to confirmatory information (Jost et al., 2022; Zmigrod et al., 2020). A more theoretical implication is that cognitive rigidity is a moderating factor of influence on beliefs and attitudes, as it shapes how people react to congruent or incongruent sources and messages. For these reasons, we see cognitive rigidity as a key variable in human-ai influence dynamics, particularly in the political context, which deserves further empirical investigation.

1.1.2 A Framework for Persuasion - Social Influence

Having outlined how beliefs are formed and updated and after expressing our special interest in dispositional factors, we now move to a concrete framework for influence between social actors. Consistent with the general process of belief formation and updating, Social influence theory (SIT) (Cialdini & Goldstein, 2004) describes how changes in beliefs, attitudes and behaviour can be caused by one social actor influencing another. It distinguishes between compliance – responses to explicit requests – and conformity, adjustment of beliefs or behaviour in light of pressure from incongruent information. In the context of the current study, we focus on conformity as this is what our design aims to produce and what we would generally expect in human-ai interactions to happen most. SIT postulates that there are two forms of influence leading to conformity - “informational” and “normative” influence (Deutsch & Gerard, 1955). Three psychological needs rest at the core of people’s susceptibility to these two forms of influence: the need to maintain an accurate mental model of the world, the need to maintain good relationships with our social in-groups, and to maintain a positive self-perception. These needs are, too, motivational factors for how people engage with sources and messages, which can be modulated by individual differences (Cialdini & Goldstein, 2004; Petty & Cacioppo, 1986). Informational influence caters towards the need for accuracy, normative influence towards the need of affiliation. Maintenance of a positive self-view is arguably involved in both forms, as both being accurate and affiliated with one’s social in-group contributes towards it. It must be noted, however, that in real-world situations these motivations often interact and are triggered simultaneously (Cialdini & Goldstein, 2004). For example, self-concept enhancement

could sometimes oppose affiliation goals (e.g. when one chooses to deviate to feel unique) or might reinforce accuracy goals (e.g. wanting to feel competent or rational).

1.2 AI-Influence on Humans

In human-ai interaction scenarios, LLMs act as a source of social influence. The question whether they exert normative influence is a little ambiguous. On the one hand, evidence exists that suggests that when acting as an interconnected group, multiple agents can exert normative influence on humans (Song et al., 2024) analogous to seminal peer-pressure studies (Asch, 1956). This is consistent with some frameworks of human-computer interaction (Nass et al., 1994). On the other hand, most current use-cases of LLM agents are one-on-one (e.g. using OpenAI's ChatGPT). Here, evidence suggests the normative influence of LLMs is limited, as participants don't feel the need to comply to a single agent to the extent they do when interacting with human individuals (Jakesch et al., 2023; Song et al., 2024). It seems fair to assume therefore that, at this point in time and as a source itself, normative influence of LLMs is limited. This might change over time. Multimodal functionalities like advanced ai voices and the anticipated combination of ai with humanoid robots increasingly anthropomorphizes these systems. Additionally, expectations of and trust in them might change over time as the technology improves and becomes more reliable, potentially making them more credible social actors or a form of authority (Sundar, 2020). Further, this is not to say that models can't already indirectly, through their messages, tap into normative influence by saying something like "most people agree that". As outlined earlier, it is hard to keep the two forms separated in real scenarios.

Less ambiguously, LLMs do already seem to have well-documented informational influence on humans. One example comes from Allan et al (2021). In their experiments, participants made decisions about visual scenes with the aid of either a human or a simulated AI partner. Results showed that participants conformed to the AI's suggestions to the same extent as to those of human partners, particularly when the AI was presented as highly credible or accurate. Notably, this conformity seemed to be a strategic trade-off. Participant's reliance on the AI increased when their own memory was weak and when the AI provided high-confidence or high-likelihood responses, thus reflecting a functional, accuracy-driven mechanism for informational influence. Additionally, this influence was unrelated to objective accuracy of the suggestion of the AI. This implies that informational influence from AI is present and regulated by much the same cognitive processes that govern memory conformity between

humans. It also suggests that people sometimes attribute unwarranted credibility to the AI, which leads to 'inappropriate conformity'.

An example specifically on political attitudes comes from Hackenburg and Margetts (2024). In their study, participants were exposed to persuasive political messages generated by GPT-4, the latest model of OpenAI, with messages either tailored to individual's demographic and political profiles (microtargeting) or generic (best message). They had to give their attitudes on a political issue before and after being exposed to the persuasive message. Both types of messages produced statistically significant shifts in political attitudes compared to a control group, in some cases increasing support for an issue by up to 12 percent. Notably, microtargeted messages did not outperform generic ones. Hackenburg and Margetts (2024) argue this suggests that the persuasive power of LLMs does not rely on personalization, but rather on the credibility and perceived quality of the information itself; an interpretation consistent with SIT, and informational influence in particular, that we share.

The findings of Hackenburg and Margetts (2024) as well as those of Allan et al. (2021) are consistent with other contemporary findings. For instance, Huang and Wang (2023) compared the persuasiveness of AI agents in varying roles (converser, curator, creator, contemplator) and different interactive and one-way communication judgement and evaluation tasks (e.g., news article consumption, customer service conversations, evaluation of government decisions) to that of human counterparts. They found that overall, AI agents were as persuasive as humans. While AI agents weren't as persuasive as humans in shaping behavioural intentions, there was no significant difference regarding perceptions of and attitudes towards the source (AI vs human), as well as actual behaviour exhibited by the human interaction partners in response to the interaction. As the tasks in their studies (persuasive messaging, debate, product evaluation) were all relating to informational influence, this interpretation is consistent with arguments made so far.

In essence, this tells us with some confidence that LLMs agents present a source of informational social influence when interacting with humans. Their influence depends on the overall perceived credibility of the agent and the quality of its message by humans, which in one-on-one conversations is driven by their need for accuracy and consistent self-perception, and subject to dispositional (e.g., cognitive rigidity) and contextual factors. Given this, a currently underpopulated part in the literature is the role that individual differences have in this influence dynamic. A lot of research currently focuses on either the general ability of persuasion or model-specific aspects like size, training or access to personalised data, when

evaluating persuasiveness (Argyle et al., 2025; Bai et al., 2023a; Cirulli et al., 2025; Hackenburg et al., 2023a, 2025; Hackenburg & Margetts, 2024; Potter et al., 2024). The human side of human-ai interaction is mostly left out so far. It is established that, between humans, dispositional factors like cognitive rigidity play an important role in influence-dynamics and related phenomena such as polarization and ideological extremism (Cialdini & Goldstein, 2004; Jost et al., 2022; Petty & Cacioppo, 1986; Zmigrod et al., 2020). LLM agents present a novel source of informational influence, one that arguably is perceived differently to human social actors (Sundar, 2008, 2020), which potentially modulates their influence compared to humans to at least some extent. Hence, one part of the effort to accurately describe how human-ai influence dynamics work is to produce empirical evidence on whether dispositional factors play the same role in human-ai interactions as they do in human-human interactions.

Another issue of the field is of a more methodological nature. Over the last two to three years, interest in using LLMs as a tool in social science research has increased drastically (Ziems et al., 2024). One challenge with this, as mentioned at the beginning, is control over the agent's responses (Sclar et al., 2024; Wang et al., 2023). Agents are sensitive to minor, even meaning-preserving changes in prompts that cause differences in output. Agents used by social science researchers are commonly prompted from a base model like GPT-4 or Llama3 via the system prompt, which is an initial prompt given to the model that is constantly kept salient throughout any sort of interactions with users. The prompting strategy many studies use is based on giving the agents instructions on a task to complete, attributes to assume and a role to play, e.g., 'you are warm and helpful' or 'you are a clinician' (Argyle et al., 2025; Bai et al., 2025; Costello et al., 2024; Hackenburg et al., 2023a, 2025; Hackenburg & Margetts, 2024). This kind of prompting uses coarsely defined roles and attributes. The model's internal interpretations of said roles and attributes can't be known by researchers, especially when using big models like Llama and GPT, as these are 'closed-weight models', meaning that their internal representations of concepts such as 'warm' or 'clinician' are not publicly accessible (and so cannot be known by researchers when using these models). The way in which a model approaches a task, e.g., whether it takes a conservative or liberal stance on a given issue, is often defined through said vague role or the task itself. For instance, Hackenburg et al (2025) instructed their models to argue in favour of specific statements when talking to participants. Said statements, through their content (e.g., "The U.K. should lift the 2-child cap on benefits, even if it encourages less well-off people to have larger families"), defined whether the model argued from a conservative or liberal perspective. Under experimental conditions this works, but it is laborious as one has to instruct the model on every topic. In practice this also makes

things complicated, as practitioners would have to anticipate all topics a model could possibly talk to people about, and give it instructions for each topic ahead of time, to ensure the model responds in the way initially intended.

If one could instead create a researcher-defined and tightly controllable baseline, an identity, to which tasks could be added and automatically executed in a way aligned with said identity, this would provide a generalizable solution to these issues. As psychologists, one metric that comes to mind for this are psychometric scales. Using them to define agent identities could open up the opportunity to use the very same scales to create, evaluate and control agents in much more nuanced ways than previously. It would potentially also be less reliant on the models internal (and often unknown) representations of a concept, as a scale provides a fixed set of attributes with specific loadings (e.g., 80/100 extrovert) to define said concept based on the definition *the scale* (e.g., big 5 personality scale), and hence the researcher, has of it. While psychometric scales are already used to examine models (Faulborn et al., 2025; Li et al., 2024; Pellert et al., 2024; Schelb et al., 2025), the idea of using them for also creating agent identities in the first place is still almost completely unexplored at this point (for the one exception found, see Huang et al., 2024).

To summarise, LLMs can in many cases be as similarly persuasive as a human when interacting with individuals. Further, the influence LLMs have on people's beliefs and attitudes operates mostly within the same theoretical frameworks that apply to human-human interaction (at least when it comes to text-based conversations). Finally, there is neither sufficient research on the role of dispositional factors, like cognitive rigidity, in said influence dynamic, nor is there an established, generalisable framework for using LLM agents as stimuli in experiments at this point.

1.3 The Current Study

Therefore, the aim of the current study is twofold. For one, we aim to contribute to the emerging literature on prompting frameworks. We propose a method we call psychometric prompting, which uses psychometric scales to create LLM agents with different identities and tests the robustness of said identities through correlated scales. This is a necessary precondition for the second aim of this study, which is to examine how carefully controlled alignment effects between participants and LLM agents interact with cognitive rigidity, and how their interaction shapes attitude-change on political issues. To test this, we created a within-subjects human-ai interaction experiment. Participants talked about four different political topics (Abortion, Welfare Benefits, Limited Government, Military & National Security) to two

LLM agents (two topics each) – one with a conservative and one with a liberal identity. Both agents were created based on psychometric prompting. Participant's attitudes towards the issues were taken once at the start of the experiment and after each conversation. Their political orientation and cognitive rigidity were also measured at the start of the experiment.

We have two research questions. The first is whether psychometric scales can be used to create stable political agent identities from LLMs. The second one is whether cognitive rigidity plays the same role in human-ai interaction as it does in human-human interaction. Informed by the theoretical background, to address these two questions we hypothesize that 1) AI agents do persuade humans based on informational social influence, also on political topics, and 2) people high in cognitive rigidity exhibit increased amount of motivated reasoning, compared to those less rigid, when interacting with LLM agents. The predictions following from this are that a) both types of agents will influence participants attitudes into their identity's direction, b) attitude change will be guided by alignment between agent and participant (more change when aligned and less when misaligned), c) there will be a significant interaction between cognitive rigidity and alignment on opinion change.

2. Methods

2.1 Participants

The study tested a total of 160 participants (80 Males, 74 Females, 4 Other, ages 18-53, M = 36, SD = 8.5) from the general public of the UK and the US. The sample contained 31 conservatives, 82 liberals and 34 centre. Education ranged from Highschool to Masters/Doctorate degrees. Participants were recruited via financial incentives (£4.50 payment) on the online crowd-sourcing platform "Testable Minds". All included participants were at least fluent in English or native speakers. 13 Participants were excluded due to a complete lack of data, leaving a total sample of N = 147. 109 out of the 147 samples were complete. This research was approved by the School of Psychology Ethics Committee at the University of Aberdeen. Informed consent was obtained by all participants. All participants were debriefed after the experiment. See appendix D for debrief materials.

2.2 Materials

2.2.1 LLM Agents

The experimental stimuli consisted of two LLM agents, one liberal and one conservative, and their conversational output. The basis for creating and controlling the LLM agents was a three-step process we call psychometric prompting. It is a scale and model agnostic framework for creating LLM agents, using prompts, based on psychometric scales. Via the item-level scores of a given scale, an identity is given to a LLM via the models' system prompt. The system prompt is an initial layer of communication, sitting 'on top' of the actual embedding space of the model, where researchers can set initial tasks, goals, traits and constrains that are salient to the model at every interaction it later has with a user. This identity is then tested using a second, correlated scale that the model answers - to see if the scores are analogous to the first scale in the expected way (e.g., an agent that was given a conservative identity scores conservative on the second scale). Based on this identity, use-case specific tasks can be given to the agents. A detailed description of our initial proof of concept of this process, using LLama, can be found in appendix A.

The LLM that was used during the experiment of this study was Llama version 3. 3B. The LLM agents were created by prompting Llama to adopt two different identities, conservative and liberal. The identities were based on two opposing answer-patterns on the 12-item social and economic conservatism scale (SEC) (Everett, 2013). The conservative identity consisted of very high and the liberal of very low conservatism scores. These answer-patterns were given to the model via the system prompt. Additionally, both agents were given the goal to convince their conversation partner of their points of view in conversations by using positive framing (See appendix A section 1 for the full prompt). Positive framing was chosen as a strategy because it is both consistent with SIT (Cialdini & Goldstein, 2004) and it mimics a common communication strategy in politics (Chong & Druckman, 2007; Hallahan, 2011; Matthes, 2012). The psychometrically prompted agents' conversational behaviour was tested informally by us prior to the experiment, to test for any abnormal behaviour – which did not occur. After the experiment, a quantitative analysis of the conversational data was done using a political classifier model, to test the agent's adherence to their identities. Results of this are provided in the results section.

To avoid order effects, the conversations on the four different issues and the two LLM agents were randomized into 8 counterbalancing conditions, where every condition contained two conversations with a liberal and two with a conservative agent, as well as both conversational directions (agent holds a positive or negative attitude towards the issue) per kind of agent. In other words, a participant of any given counterbalancing condition always spoke to both agents twice, once about a topic the agent had a positive attitude towards and once about a topic the agent had a negative attitude towards.

2.2.2 HAX Custom Application

HAX is a custom platform for human-ai interaction experiments developed in the School of Psychology of the University of Aberdeen. It allows researchers to create experiments that contain questionnaires and stimuli, one of which can be conversations with psychometrically prompted LLM agents. It was used in the current study to administer all questionnaires and the conversations between humans and agents. As this application is still under active development, please contact the researchers for further information or access.

2.2.3 Pythax Custom Application

Pythax is another custom application developed for this and other human-ai interaction studies. It allows for a streamlined process of psychometric prompting using model such as GPT-4 and Llama 3.3B. That is, it provides a user interface to systematically define and validate agent identities, based on said models, using psychometric scales, as well as to compare the transfer performance of different identities against one another. It further enables LLM-based text analysis, such as political classification, through locally hosted versions of models from platforms like OpenAI and huggingface. The results in appendix A, as well as the political classifications agent messages in the results section, are produced with Pythax. As this application is also still under active development, please contact the researchers for further information or access.

2.2.4 Political Classifier

The classifier model “political-leaning-deberta-large” by Matous Volf (Volf & Simko, 2025) was used to post-hoc analyse the transcripts of conversations between participants and LLM agents. It is an open source model on the platform huggingface. The version used in this

experiment did not contain the initial pre-classifier, which distinguishes between political and non-political messages, for technical reasons.

2.3 Design

This study was a within-subjects design with the independent variable of alignment. Alignment had two conditions, aligned and misaligned. Participants cognitive rigidity scores served as a second, continuous predictor. The dependent variable was the change in participant's attitude scores, calculated as the difference of participant's initial SEC scores and their post-conversation scores on the issues they talked about, which were 4 out of the 12 SEC issues.

2.4 Procedure

Participants started the experiment on the crowd-sourcing platform Testable Minds, where they filled in a consent form and some demographic data (age, sex, education, nationality, English proficiency). They were then directed, via a link, to the experimental platform HAX. Once the experiment started, they were instructed to first answer three different questionnaires. The first questionnaire was an 11-point left-right self-placement scale (Kroh, 2007), which measured explicit political leaning on a left-right dimension. The second questionnaire was the 12-item Social and Economic Conservatism Scale (Everett, 2013) (SEC), which measured participant's conservatism and served as a more nuanced metric of political orientation. The third one was the Cognitive Flexibility Inventory (Dennis & Vander Wal, 2010), which captured participants cognitive flexibility. Participants were then instructed that they would have a conversation that would last 5 minutes and would automatically end after that, and to start the conversation by sending 'Hello' in the chat. This instruction was repeatedly given to participants before each conversation. They went through four conversations, two with each type of agent, and always on the same four issues (Abortion, Welfare Benefits, Military and Security, Limited Government). The order of the issues depended on their counterbalancing condition. After each conversation, they were instructed to fill in a short questionnaire asking a single issue of the SEC scale, which was always the issue from the conversation immediately prior. Their answers to the questionnaires were scored as prescribed by the authors (For a list of all scales, see appendix 3).

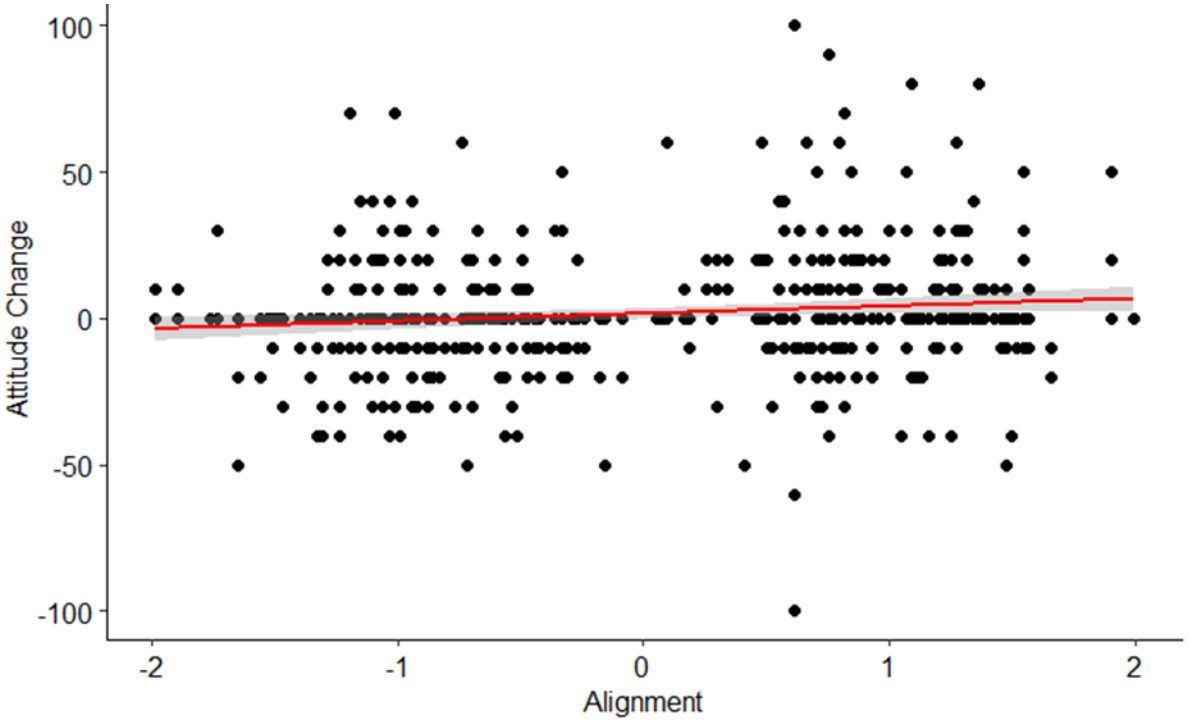
3.Results

3.1 Agent Stability

Our first prediction was that both agents would influence participant's attitudes into the direction of their identities. This would be reflected in the data through participants changing their attitude *into the direction of the agent's political leaning* after interacting with them. This pattern would support the notion that we successfully created alignment effects between participants and agents through stable agent identities, as the latter is a necessary precondition for the former.

***Figure* 1**

Relationship Between Alignment and Attitude Change



Note. Relationship between continuous version of alignment and attitude change

Figure 1 shows the relationship between the direction of attitude change of participants (positive or negative) and a variable called alignment. Both attitude change and this continuous version, different from a categorical version used further down, of alignment are based on the SEC. The SEC measures conservatism so that higher scores on items and overall reflect higher conservatism. Positive attitude-change scores therefore reflect the participant changing their attitude towards a more conservative stance on issues, negative scores reflect a change towards a more liberal attitude - compared to their initial SEC scores. The 'alignment' variable here was created by subtracting the initial SEC scores of participants from those of the LLM

agents (which had been given to them in the first step of the psychometric prompting process). Hence, On the x-axis, 0 reflects a continuous form of alignment defined as both agent and participant scoring the same on the SEC. However, here it is primarily used to show whether agents attract opinion change in their direction. Positive values on the x-axis reflect the agent being more conservative than the participant, negative scores that the agent was more liberal than the participant. Therefore, values in the top-right and bottom-left quadrants of figure 4 reflect participants changing their attitude on a given issue *towards the bias of the agent*, successful persuasion in other words. The regression line shows a slight trend into that direction.

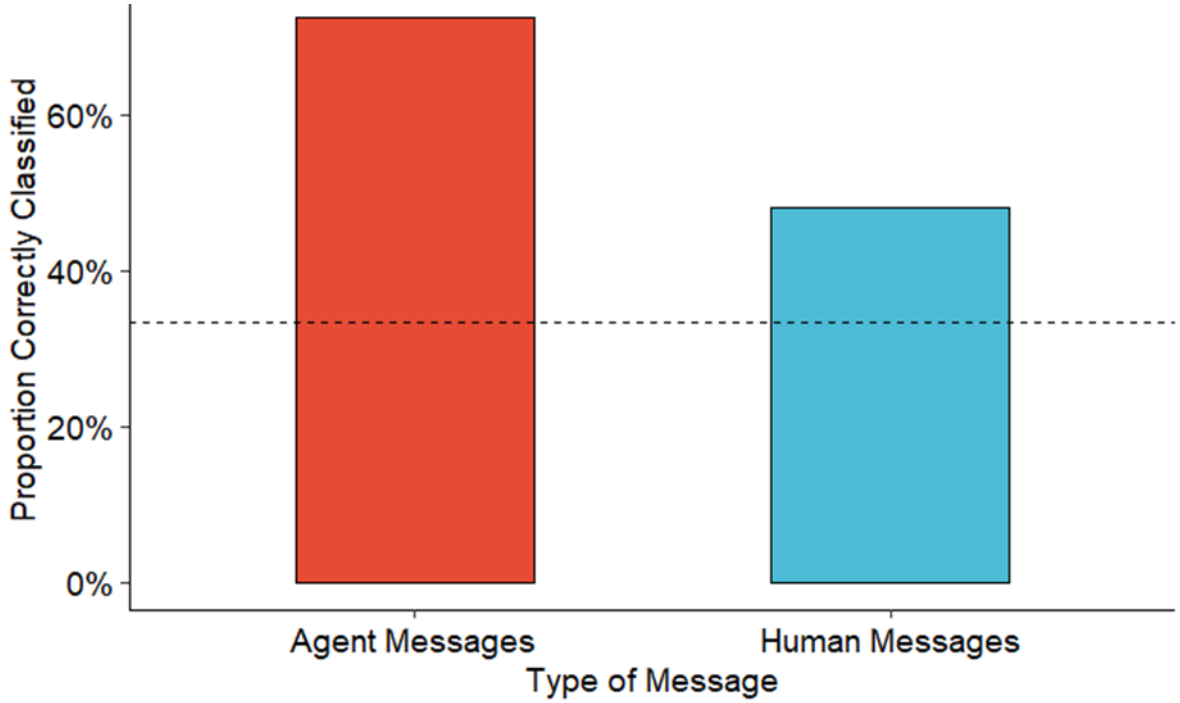
To further test this, via the lme4 (Bates et al., 2015) (for the model itself) and lmerTest (Kuznetsova et al., 2017) (for p-values calculated using the Satterthwaite method) packages in R, a mixed effects model testing for the effects of continuous alignment and cognitive rigidity on attitude change was run on the data. The model included all fixed effects and the following random structure: random intercept for participants. continuous predictors were centred and standardised. The fitting strategy followed the maximum approach, simplifying via least explained variance. Due to convergence issues, only the fixed effects structure will be reported. Repeating the calculations as a multiple linear regression, to control for increased type1 error rates in models with only intercepts (Brauer & Curtin, 2018), revealed identical results - justifying this decision. The analysis showed a significant effect of continuous alignment ($t = 2.73$, $p = .007$) on opinion change. There was no significant effect of cognitive rigidity ($t = 1.18$, $p = 0.24$) and no interaction between alignment and cognitive rigidity ($t = 1.61$, $p = 0.11$). This suggests that the attitude change that did occur in participants was significantly into the direction of the agent's bias, and that we did create two different identities.

A second form of analysis further tested for stable agent identities using the off-the-shelf classifier model "political-leaning-deberta-large" by Matous Volf, available on the website huggingface.co. The classifier is a LLM trained to label political messages, mostly tweets, as liberal, centre or conservative. It was integrated into one of our custom applications, Pythax, and used on the conversational transcripts from participants conversations with LLM agents. The goal was to test whether the conservative and liberal agents reliably sent politically conservative and liberal messages respectively. The classifier usually consists of two 'sub-classifiers', the first classifying the message as political or not, the second labelling the political

leaning. In the current study, only the latter was used due to technical issues with setting up the former.

Figure 2

Classification Accuracy: Agents vs Humans



Note. Differences in correctly classified proportions of agent and human messages relative to total messages. Black dotted line represents chance (33%) level.

Figure 2 shows the proportion of correctly classified messages, relative to total messages (3314 for agents, 3335 for humans) across issues and alignment conditions, for LLM agents and participants respectively. This proportion was 72.4 % for agent and 48% for participant messages. Two one-sample t-tests were conducted to test whether these proportions differed significantly from chance (33%). Results revealed that both agent ($t(3313) = 50.28$, $p < .001$) and human ($t(3334) = 16.92$, $p < .001$) messages were correctly classified significantly above chance. This lends further support to psychometric prompting as a reliable way to create alignment conditions between agents and participants, as an incomplete off-the-shelf classifier model, not specifically trained on the data format present in the experiment, correctly detected agent bias in 72% of all messages.

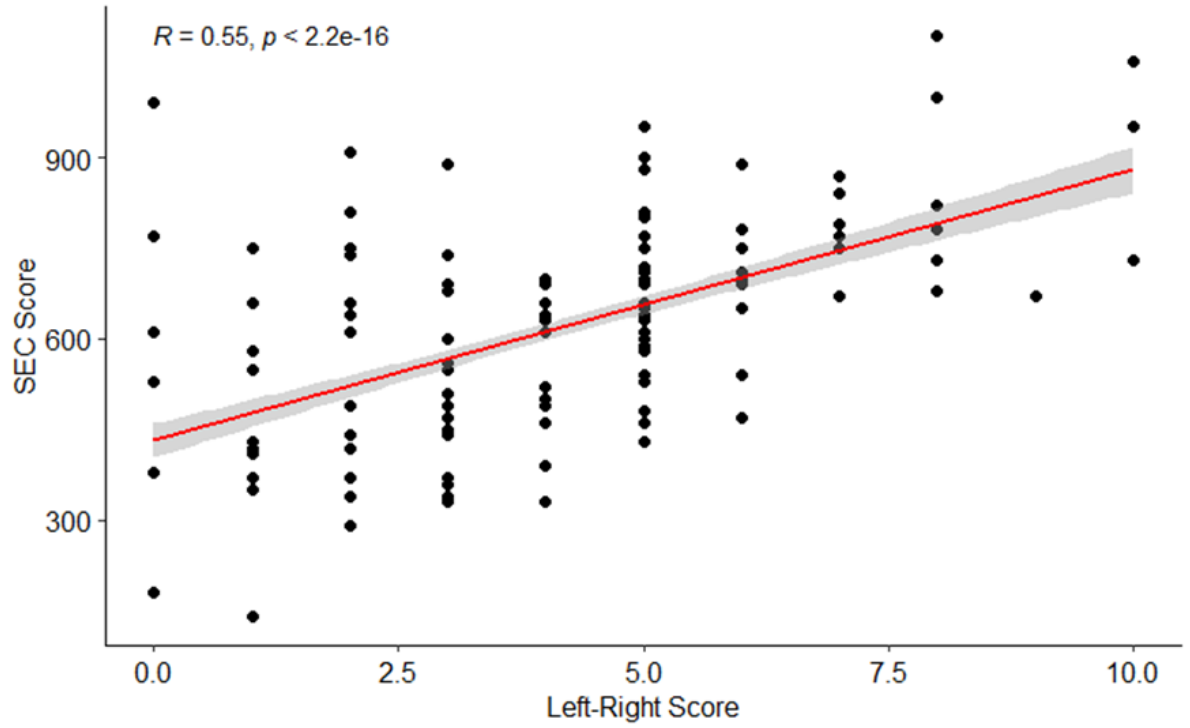
3.2 Effects of Alignment and Rigidity

The second prediction was that alignment will guide the influence agents have on participants attitudes. The third prediction was that there would be a significant interaction

between cognitive rigidity and alignment. Therefore, we were predominantly interested in the main effect of alignment and the interaction between alignment and cognitive rigidity. To test this, it was decided in advance to use a categorical version of alignment between participants and agents, alongside the continuous measure of cognitive rigidity, to predict attitude change. This categorical version of alignment was based on the agent's identity and the participant's score on the self-placement left-right measure (e.g. if participant LR score <5 & agent == liberal = aligned).

Figure 3

Relationship Between Participant's Left-Right and SEC Scores

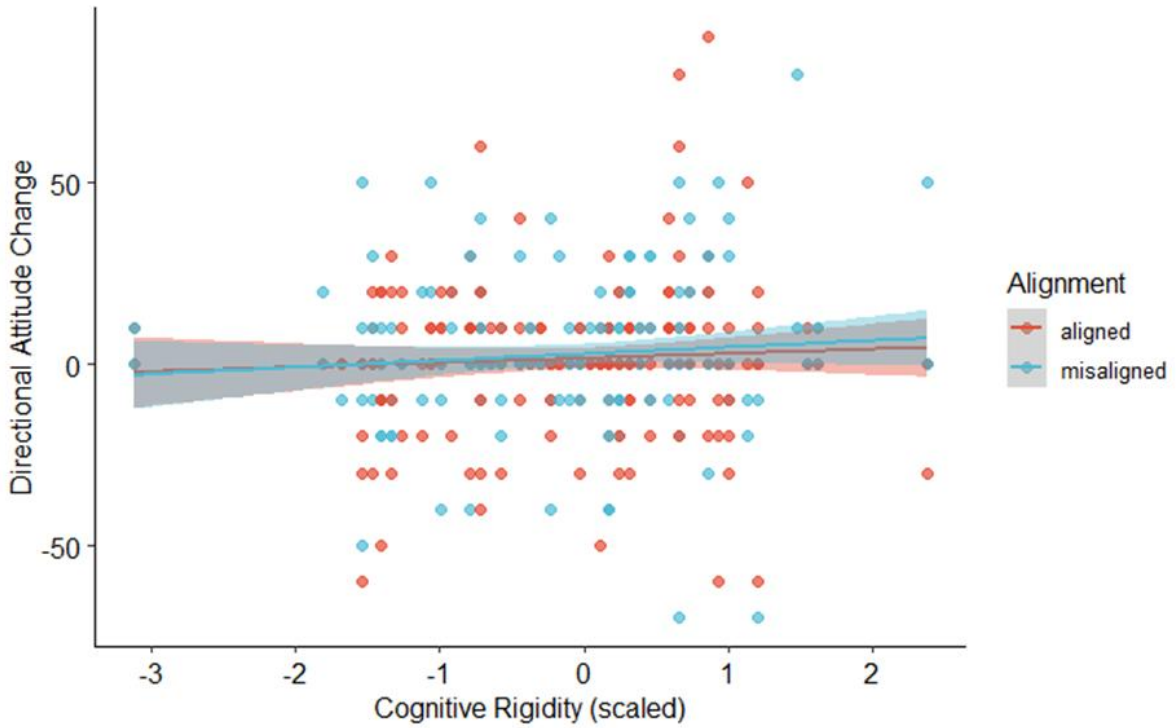


Note: Relationship between participant's Left-Right scores and their initial SEC scores

Figure 3 shows the relationships between participant's LR and SEC scores. A Pearson correlation coefficient reveals a moderately strong correlation of the two (r (489) = .55 , p < .01). This positive correlation between both measures of political orientation is worth noting as it shows that, while not perfect, there is validity in the categorisation. Further, the dependent variable attitude change was slightly modified by multiplying it by -1 or +1 depending on which agent participants talked to in a given conversation. Through this, positive attitude-change scores (as shown in figure 5) reflect change towards the agent, negative indicates a shift away from the agent leaning.

Figure 4

Attitude Change by Cognitive Rigidity and Categorical Alignment



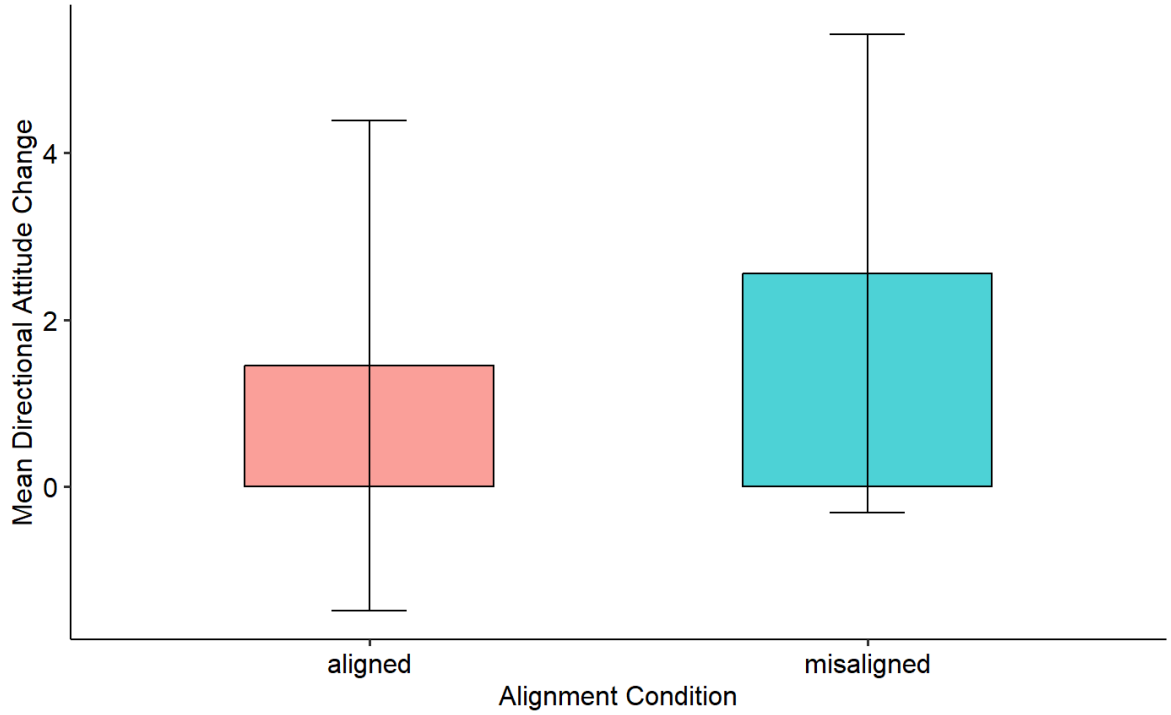
Note. Relationship between directional attitude change, categorical alignment and cognitive rigidity.

Figure 4 shows our main analysis, the relationship between said directional attitude change and categorical alignment, as described above, and cognitive rigidity. No clear effect of alignment on opinion change is shown. Further, no clear effect of cognitive rigidity can be deduced from this figure.

A mixed effects model (again via lme4 and lmerTest) was applied to the data, using categorical alignment and cognitive rigidity as predictors and directional instead of 'nondirectional' attitude change as dependent variable. The model included all fixed effects and the following random structure: random intercept for participants. Categorical predictors were regression style coded (-0.5, 0.5) and continuous predictors were centred and standardised. The fitting strategy followed the maximum approach, simplifying via least explained variance. Due to convergence issues, only the fixed effect's structure will be reported. Repeating the calculations as a multiple linear regression, to control for increased type1 error rates in models with only intercepts (Brauer & Curtin, 2018), revealed identical results - justifying this decision. The analysis showed no significant effect of categorical alignment ($t = 0.55$, $p = .57$) on opinion change. There further was no significant effect of cognitive rigidity ($t = 1.439$, $p = .15$) and no significant interaction between categorical alignment and cognitive rigidity ($t = 0.283$, $p = .78$. This suggests that, overall, agents weren't persuasive, and cognitive rigidity did not moderate what little influence agents had on participant's attitude.

Figure 5

Mean Directional Opinion Change by Alignment Condition



Note. Shows the mean opinion change when splitting data by alignment. Error bars show 95% confidence intervals.

Figure 5 shows the mean directional attitude change of the aligned and misaligned condition. Together with the results of the mixed effects model, this figure with widely overlapping confidence intervals underlines the non-significant effect of alignment on directional attitude change.

4. Discussion

Our results showed, through a mixed effects model predicting attitude change from continuous alignment and cognitive rigidity, a significant effect of continuous alignment on attitude change. This analysis addressed our first prediction, that both agents would influence participants attitudes into the direction of their identity. So did the results from our classifier model. The main analysis, a second mixed effects model predicting directional attitude change from categorical alignment and cognitive rigidity, found no significant effect of categorical alignment on attitude change. It further found no interaction between categorical alignment and cognitive rigidity. The mixed effects model analysis addressed predictions two and three, that alignment would guide the influence agents have on participants attitude and that cognitive rigidity would significantly interact with alignment.

4.1 Psychometric Prompting

Our first research question, whether psychometric scales can be used to create stable agent identities, is supported by the findings. The influence that agents did have on participants attitudes was into the direction of the agent's identities - revealed through the significant effect of our continuous 'alignment' predictor on attitude change. This could only happen if we did, via psychometric prompting, create reliably different agent identities. This interpretation is further supported by the results of the conversational classifier model we used (Volf & Simko, 2025), which accurately classified 72% of agent's messages as liberal or conservative. While 72% is not a perfect match, given the exploratory component of this framework this still gives us confidence that psychometric prompting as a strategy worked at a basic level. The classifier model we used was an 'off-the-shelf' open-source model that was trained mainly on tweets, therefore standalone statements, not on conversational text data. It was further missing the initial classifier distinguishing between political and non-political messages. This potentially could have led to non-political messages, like those initiating the conversation or when conversations deviated from the topic, being classified too. Finally, even on its original training data, the classifier doesn't reach 100 but only 87% correct classifications on average (Volf & Simko, 2025). Even at the 72% rate, conservative and liberal agents at least mostly behaved in line with their identity, making psychometric prompting seem like a promising approach to further evaluate.

Our findings are also consistent with the single other study we are aware of, at the time of writing, to also use psychometric scales (in their case the big five inventory 2 (BFI2)) to create agents (M. Huang et al., 2024). They too report successfully creating agents adhering to the psychometric profile they were given. They measured this for example through correlations between the BFI2 scale and the mini markers scale (a different form of Big five test). Said correlations were 0.664 on average, mimicking that of humans. This is analogous to the transfer scale step in our psychometric prompting approach. Notably, they also used a promising approach for evaluating agent identities via decision-making behaviour. They presented agents with decision-making tasks that are known to reflected personality traits (e.g., high-risk decision making indicates openness and extraversion) in humans to test whether an agent with a given identity would behave in the same way humans with an analogous identity usually do. There are behavioural measures suggested by research to be predictive of political leaning (Claessens et al., 2023; Shook & Fazio, 2009), and while not all seem to be equally

implementable, this could be a fruitful approach to explore in future research to improve behavioural evaluation of agents.

4.2 LLM Influence on Participant's Attitudes

While our first hypothesis was corroborated by the findings, our second one, that attitude change would be guided by alignment, was not. Agents in this experiment did not have the expected informational influence on participant's attitudes towards political issues, as demonstrated by the absence of a significant effect of *categorical* alignment on attitude change. This makes our findings inconsistent with a lot of literature examining the influence of AI on humans in a broader context (Allan et al., 2021; G. Huang & Wang, 2023), on general issue-related attitudes (Costello et al., 2024; Xu et al., 2025), and attitudes on political issues in particular (Argyle et al., 2025; Bai et al., 2023b, 2025; Hackenburg et al., 2023b, 2025; Hackenburg & Margetts, 2024; Potter et al., 2024; Salvi et al., 2024).

This could have different reasons. In general, effect sizes in persuasion studies tend to be relatively small (Coppock et al., 2020). This also seems to be the case in human-ai interaction scenarios. For instance, Hackenburg et al (2025) and Argyle (2025), while not reporting Cohen's d values, achieved between 2 and 12 percentage points attitude-shifts in their studies, with most being closer to 5 pp (12 was reached with specifically post-trained models). They also had significantly bigger sample sizes compared to us (e.g., N=76000, N = 3600), making it easier to detect even small effects. This is the same throughout comparable studies (Argyle et al., 2025; Bai et al., 2025; Hackenburg et al., 2023b; Hackenburg & Margetts, 2024). Therefore, at a sample-size of 147, insignificant findings aren't completely unexpected. Our study further suffered from data attrition, most likely due to technical issues with our custom platform HAX, causing people to drop out mid-experiment and leaving us with only 109 complete datasets. More data was lost in the main analysis through categorization of alignment as a predictor, as only participants categorized as conservative or liberal were included (34 neutral participants were excluded). The number of complete datasets of liberals and conservatives combined was 86. An a-priori power analysis arrived at a sample size of N = 150 to reveal small to moderate effects using our mixed effects model, effectively leaving this study underpowered. The technical platform issues, as well as the form of alignment predictor, will be addressed in future analyses to avoid sample size and attrition as potential reasons for incongruent findings.

Cognitive Rigidity as a Moderator of LLM-Social Influence in Political Contexts

Turning to theory, another possible explanation for our findings could be that LLM agents as a source weren't appealing, and their arguments as a type of information less relevant to both aligned and misaligned participants, because of the experimental conditions. As mentioned earlier, information can be processed more peripheral and more elaborative. Depending on what kind of processing people engage in, different aspects of messages are persuasive. The quality of the argument is an aspect that matters most when people engage deeply with the information. If they don't, the perception of the source plays a more important role (Cialdini & Goldstein, 2004; Petty & Cacioppo, 1986) . In our study, the conversations were restricted to 5 minutes. Additionally, the agents' output was not constrained to a particular length, making some of their arguments longer than others. Long arguments and limited time could have reduced people's processing capacity, potentially leading to reliance on peripheral information cues (source) rather than paying attention to the quality of the arguments (Evans & Stanovich, 2013; Kruglanski & Gigerenzer, 2011; Petty & Cacioppo, 1986). Additionally, we did not measure people's interest for the issues they were talking about or politics in general. Personal relevance also increases (high) or decreases (low) the likelihood of more elaborate processing (Petty & Cacioppo, 1986). This could have pushed for peripheral processing even more. If participants therefore were paying less attention to message quality and focused more on the source instead, the credibility of the LLM agents as a source of information on political issues would have played a decisive role in whether influence was had or not. As AI can potentially be perceived as manipulative and untrustworthy depending on the context (Sundar, 2008, 2020), and our agents were designed to push participants into one direction, this could have led to a categorical rejection of the AI's arguments, based on the combination of contextual factors.

This seems plausible at first glance, as there are differences in experimental design between our study and literature with significant results. For instance, Hackenburg and Margetts (2024), as well as Bai et al (2023), used a single message to persuade participants and took care to leave them unaware that they were talking to an LLM as, according to them and consistent with our reasoning, evidence suggests that awareness of talking to AI reduces message persuasiveness (Hackenburg & Margetts, 2024). Salvi et al (2024) used a debate format (opening, rebuttal, conclusion) with more time (10 minutes per debate) and restricted message length (1-2 sentences). In these studies people had a) more time to engage with information and b) weren't explicitly aware of the synthetic nature of the information they encountered. This could have led to the quality of argument being more important, or the

source not being a limiting factor for persuasion compared to our study, leading to significant findings.

However, our results are also inconsistent with research where participants *did* have interactive conversations with LLM agents of reasonably similar format, on political issues, while being aware of talking to LLM agents (Argyle et al., 2025; Hackenburg et al., 2025). Furthermore, evidence suggests that interactive conversations are more persuasive than static messages (Hackenburg et al., 2025), which should make the agents of the current study more persuasive, rather than less, compared to studies using single messages (but see also Argyle et al for contradicting results (2025)). Additionally, information-density is apparently a key factor boosting persuasiveness in human-ai interaction (Hackenburg et al., 2025), suggesting participants do pay attention to the quality of the argument in interactive conversations with LLMs akin to the ones in our study. This leaves us with the assumption that, even considering the theoretical perspective, aligned and misaligned agents could have had the expected impact on participant's attitudes under our experimental conditions. That is, assuming they behaved sufficiently in accordance with their identities – far right and far left-leaning.

A possible explanation for why this might have not been the case is the agent's prompting and their conversational output. It is unlikely that the type of model made a big difference on how the models behaved or how persuasive they were, as evidence for persuasiveness has been produced across different models, including the llama 3 model family (Hackenburg et al., 2025). Regarding the prompting, our model's system prompt included both the psychometric identity of the agents and the positive framing persuasion strategy (see appendix A section 1). No comparable study testing political attitude change through LLM agents used psychometric scales as part of their prompting. Almost all of them did used some form of a 'persuasion strategy' via the system prompt however (Argyle et al., 2025; Hackenburg et al., 2023b; Salvi et al., 2024). While some arguably weren't completely different from ours, they certainly weren't identical. For instance, Hackenburg et al (2025) tested persuasiveness of llama by giving the model, among others, an 'information' strategy. It instructed the model to persuade through "providing information, evidence, and context, clearly communicating complex facts and making them accessible" (pp. 69 in supplementary materials). The political leaning was given to the model via the issue, by telling the model to argue in favour of a certain statement regarding the issue that was either conservative or liberal (Hackenburg et al., 2025). Our model was instructed to persuade by "consistently presenting one side of an issue in terms of its gains, benefits, and potential positive outcomes". The stance on the issue the agent took was dependent on the identity (liberal or conservative). As mentioned earlier, models are

sensitive to various aspects of changing prompts, even if they are meaning-preserving (Elazar et al., 2021; Röttger et al., 2024; Sclar et al., 2024; Wang et al., 2023). Therefore, differences in the persuasion strategy itself, as well as the inclusion of the psychometric identity, could have played a part in leading to difference in conversational output and, potentially, incongruent findings.

Having both psychometric identity and persuasion strategy in the system prompt was necessary given the nature of the experiment (persuasion on political issues). But it is also a limitation in this case, as there was no control group talking to agents prompted with no or only minimal task instructions. This makes it impossible, at this point, to say whether the identity itself, the prompting strategy or a combination of both guided conversational output and therefore persuasiveness. This should be tackled in future studies by systematically testing the identity part of the prompt more isolated, and by comparing the strategy part to that of other studies (e.g. those in Hackenburg et al (2025)).

Another limitation of this and most other studies on human-AI interaction involving interactive conversations, is that it is challenging to measure the agent's conversational performance in meaningful ways. We examined whether our agents adhered to their identity in the most fundamental way (whether they presented liberal or conservative stances on issues). This doesn't tell us *how far* right or left they behaved, however. For instance, our persuasion strategy also included the notion to "subtly influence" and to "not directly attack the opposing side". This might have made the model "too understanding" to be persuasive, even while being conservative or liberal respectively. In the current study, we did not have the tools to analyse the conversational output of our agents any further. Methods to evaluate conversational output employed by other studies include for example checking for informational density (number of fact-checkable claims) using LLMs, and checking for factuality using both LLMs and human fact-checkers (Hackenburg et al., 2025). Another example is visualising semantic differences of messages by turning them into vectors and embedding them in a high-dimensional vector space (Argyle et al., 2025). The former is useful to test certain aspects of a persuasion strategy but is very resource intensive. The latter could potentially be used in our case to test how different messages from different identities are. This too comes with limitations, however, as axis labels aren't interpretable with this method and therefore won't allow to tell in what exact way messages are different. Future research should aim to find and test more ways to analyse conversational output of LLMs in more detail, as they seemingly present, alongside the previously mentioned decision-making tasks, the best measure of stability for LLM identities we are currently aware of.

4.3 The Role of Cognitive Rigidity

Our final prediction, that we would find a significant interaction between cognitive rigidity and alignment, wasn't supported either. There are plausible reasons for this. The fact that there wasn't a significant effect of categorical alignment on participant's attitudes is in itself a valid explanation for the absence of a significant interaction between cognitive rigidity and categorical alignment, and therefore any moderating role of cognitive rigidity on informational influence. Thus, we see the absence of significant effects of cognitive rigidity as a direct consequence of the possible reasons and limitations discussed above in relation to informational influence guided by alignment. Informed by literature, since we measured cognitive rigidity through a self-report scale, we would expect this trait to be more prevalent in right-leaning participants, reflected by a positive correlation between participants SEC and cognitive rigidity scores. We would also expect the moderating effect of cognitive rigidity to be higher in the right-leaning part of the sample. This analysis would be based on participant's political leaning which, given the current sample's size, unbalanced political nature (31 conservatives, 81 liberals, 34 neutral), as well as the mentioned attrition rate, wouldn't be promising in the current study. This should be addressed by future analyses with a bigger, more complete and politically balanced sample.

In light of what was discussed, we take away three things from this study. The first one is that psychometric prompting is a promising framework that was supported, even under otherwise not ideal conditions, by the results. It seems possible to define an identity that transfers to at least a reasonable degree into conversational behaviour of the agent, which offers potential benefits compared to currently used prompting methods (e.g., generalisability, control). This needs further validation, mostly via tools for conversational output analysis, to examine how far the control of conversational behaviour one gains, compared to other prompting strategies, can go. The second takeaway is that while our findings are incongruent with a lot of the existing literature on LLM persuasion on political issues, there seem to be many plausible reasons for this that don't make this a clearly contradictive study. These reasons mostly seem to be of a methodological nature (sample-size, wording of the persuasion strategy part of the prompt, tools for analysing conversational output) and can be addressed by future research. This should be done before a definitive conclusion on informational influence of LLMs on political attitudes, or the extent of efficacy of psychometric prompting from *our* perspective, can be reached. The third and final takeaway is that these limitations made it effectively

impossible to properly examine the role of cognitive rigidity as a moderating factor, leaving this as an unknown. There is still reason to assume that it has the role suggested by us and understanding the human factors in human-ai influence dynamics remains an important point in our opinion. Once the discussed limitations are addressed, another attempt can be taken to reveal the role of this dispositional factor in human-ai influence dynamics.

Conclusion

This study set out to lay the foundation for psychometric prompting as a framework for using LLMs in psychology experiments in political contexts. It further set out to apply this to answer the psychological research question whether the dispositional factor cognitive rigidity has the same moderating impact on the social influence of LLMs that it seems to have on that of humans. We found promising results for psychometric prompting. Although much more work needs to be done to further corroborate it. We did not find significant persuasive power of LLM agents, and no significant impact of cognitive rigidity. However, there seem to be possible experimental reasons for it, which need to be addressed by future research before any conclusion can be reached.

Acknowledgements

I want to thank my supervisor, Dr Kevin Allan, for the never-ending support during this project. I learned a lot over the course of this project, and much of it was thanks to him.

References

Allan, K., Azcona, J., Sripada, Y., Leontidis, G., Sutherland, C., Phillips, L. H., & Martin, D. (2024). *Stereotypical Bias Amplification, and Reversal, in an Experimental Model of Human Interaction with Generative AI*. https://doi.org/10.31234/osf.io/r7vf5

Allan, K., Oren, N., Hutchison, J., & Martin, D. (2021). In search of a Goldilocks zone for credible AI. *Scientific Reports*, *11*(1), 13687. https://doi.org/10.1038/s41598-021-93109-8

Argyle, L. P., Busby, E. C., Gubler, J. R., Lyman, A., Olcott, J., Pond, J., & Wingate, D. (2025). Testing theories of political persuasion using AI. *Proceedings of the National Academy of Sciences*, *122*(18), e2412815122. https://doi.org/10.1073/pnas.2412815122

Asch, S. E. (1956). Studies of independence and conformity: I. A minority of one against a unanimous majority. *Psychological Monographs: General and Applied*, *70*(9), 1–70. https://doi.org/10.1037/h0093718

Bai, H., Voelkel, J., Eichstaedt, J., & Willer, R. (2023a). *Artificial Intelligence Can Persuade Humans on Political Issues*. Springer Science and Business Media LLC. https://doi.org/10.21203/rs.3.rs-3238396/v1

Bai, H., Voelkel, J., Eichstaedt, J., & Willer, R. (2023b). *Artificial Intelligence Can Persuade Humans on Political Issues*. https://doi.org/10.21203/rs.3.rs-3238396/v1

Bai, H., Voelkel, J. G., Muldowney, S., Eichstaedt, J. C., & Willer, R. (2025). LLM-generated messages can persuade humans on policy issues. *Nature Communications*, *16*(1), 6037. https://doi.org/10.1038/s41467-025-61345-5

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using **lme4**. *Journal of Statistical Software*, *67*(1). https://doi.org/10.18637/jss.v067.i01

Bengio, Y., Hinton, G., Yao, A., Song, D., Abbeel, P., Darrell, T., Harari, Y. N., Zhang, Y.-Q., Xue, L., Shalev-Shwartz, S., Hadfield, G., Clune, J., Maharaj, T., Hutter, F., Baydin, A. G., McIlraith, S., Gao, Q., Acharya, A., Krueger, D., … Mindermann, S. (2024). Managing extreme AI risks amid rapid progress. *Science*, *384*(6698), 842–845. https://doi.org/10.1126/science.adn0117

Brauer, M., & Curtin, J. J. (2018). Linear mixed-effects models and the analysis of nonindependent data: A unified framework to analyze categorical and continuous independent variables that vary within-subjects and/or within-items. *Psychological Methods*, *23*(3), 389–411. https://doi.org/10.1037/met0000159

Breum, S. M., Egdal, D. V., Gram Mortensen, V., Møller, A. G., & Aiello, L. M. (2024). The Persuasive Power of Large Language Models. *Proceedings of the International AAAI Conference on Web and Social Media*, *18*, 152–163. https://doi.org/10.1609/icwsm.v18i1.31304

Burtell, M., & Woodside, T. (2023). *Artificial Influence: An Analysis Of AI-Driven Persuasion* (arXiv:2303.08721). arXiv. https://doi.org/10.48550/arXiv.2303.08721

Chong, D., & Druckman, J. N. (2007). Framing Theory. *Annual Review of Political Science*, *10*(1), 103–126. https://doi.org/10.1146/annurev.polisci.10.072805.103054

Cialdini, R. B., & Goldstein, N. J. (2004). Social Influence: Compliance and Conformity. *Annual Review of Psychology*, *55*(1), 591–621. https://doi.org/10.1146/annurev.psych.55.090902.142015

Cirulli, D., Cimini, G., & Palermo, G. (2025). *How Large Language Models play humans in online conversations: A simulated study of the 2016 US politics on Reddit* (arXiv:2506.21620). arXiv. https://doi.org/10.48550/arXiv.2506.21620

Claessens, S., Sibley, C. G., Chaudhuri, A., & Atkinson, Q. D. (2023). Cooperative and conformist behavioural preferences predict the dual dimensions of political ideology. *Scientific Reports*, *13*(1). https://doi.org/10.1038/s41598-023-31721-6

Coppock, A., Hill, S. J., & Vavreck, L. (2020). The small effects of political advertising are small regardless of context, message, sender, or receiver: Evidence from 59 real-time randomized experiments. *Science Advances*, *6*(36). https://doi.org/10.1126/sciadv.abc4046

Costello, T. H., Pennycook, G., & Rand, D. G. (2024). Durably reducing conspiracy beliefs through dialogues with AI. *Science*, *385*(6714). https://doi.org/10.1126/science.adq1814

Crawford, K. (2021). *The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale University Press.

Dennis, J. P., & Vander Wal, J. S. (2010). The Cognitive Flexibility Inventory: Instrument Development and Estimates of Reliability and Validity. *Cognitive Therapy and Research*, *34*(3), 241–253. https://doi.org/10.1007/s10608-009-9276-4

Deutsch, M., & Gerard, H. B. (1955). A study of normative and informational social influences upon individual judgment. *The Journal of Abnormal and Social Psychology*, *51*(3), 629–636. https://doi.org/10.1037/h0046408

Dillion, D., Mondal, D., Tandon, N., & Gray, K. (2025). AI language model rivals expert ethicist in perceived moral expertise. *Scientific Reports*, *15*(1), 4084. https://doi.org/10.1038/s41598-025-86510-0

Elazar, Y., Kassner, N., Ravfogel, S., Ravichander, A., Hovy, E., Schütze, H., & Goldberg, Y. (2021). Measuring and Improving Consistency in Pretrained Language Models. *Transactions of the Association for Computational Linguistics*, *9*, 1012–1031. https://doi.org/10.1162/tacl_a_00410

Ellemers, N., Spears, R., & Doosje, B. (2001). *SELF AND SOCIAL IDENTITY*.

El-Sayed, S., Akbulut, C., McCroskery, A., Keeling, G., Kenton, Z., Jalan, Z., Marchal, N., Manzini, A., Shevlane, T., Vallor, S., Susser, D., Franklin, M., Bridgers, S., Law, H., Rahtz, M., Shanahan, M., Tessler, M. H., Douillard, A., Everitt, T., & Brown, S. (2024). *A Mechanism-Based Approach to Mitigating Harms from Persuasive Generative AI* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2404.15058

Erikson, R. S., & Tedin, K. L. (2023). *American Public Opinion: Its Origins, Content, and Impact* (11th ed.). Routledge. https://doi.org/10.4324/9781003326847

Evans, J. St. B. T., & Stanovich, K. E. (2013). Dual-Process Theories of Higher Cognition: Advancing the Debate. *Perspectives on Psychological Science*, *8*(3), 223–241. https://doi.org/10.1177/1745691612460685

Everett, J. A. C. (2013). The 12 Item Social and Economic Conservatism Scale (SECS). *PLoS ONE*, *8*(12), e82131. https://doi.org/10.1371/journal.pone.0082131

Floridi, L. (2024). Hypersuasion – On AI's Persuasive Power and How to Deal with It. *Philosophy & Technology*, *37*(2). https://doi.org/10.1007/s13347-024-00756-6

Fransen, M. L., Smit, E. G., & Verlegh, P. W. J. (2015). Strategies and motives for resistance to persuasion: An integrative framework. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.01201

Grace, K., Stewart, H., Sandkühler, J. F., Thomas, S., Weinstein-Raun, B., & Brauner, J. (2024). *Thousands of AI Authors on the Future of AI* (arXiv:2401.02843). arXiv. https://doi.org/10.48550/arXiv.2401.02843

Grant, D. A., & Berg, E. (1948). A behavioral analysis of degree of reinforcement and ease of shifting to new responses in a Weigl-type card-sorting problem. *Journal of Experimental Psychology*, *38*(4), 404–411. https://doi.org/10.1037/h0059831

Greenberg, J., & Jonas, E. (2003). Psychological motives and political orientation--The left, the right, and the rigid: Comment on Jost et al. (2003). *Psychological Bulletin*, *129*(3), 376–382. https://doi.org/10.1037/0033-2909.129.3.376

Hackenburg, K., Ibrahim, L., Tappin, B. M., & Tsakiris, M. (2023a). *Comparing the persuasiveness of role-playing large language models and human experts on polarized U.S. political issues*. https://doi.org/10.31219/osf.io/ey8db

Hackenburg, K., Ibrahim, L., Tappin, B. M., & Tsakiris, M. (2023b). *Comparing the persuasiveness of role-playing large language models and human experts on polarized U.S. political issues*. https://doi.org/10.31219/osf.io/ey8db

Hackenburg, K., & Margetts, H. (2024). Evaluating the persuasive influence of political microtargeting with large language models. *Proceedings of the National Academy of Sciences*, *121*(24), e2403116121. https://doi.org/10.1073/pnas.2403116121

Hackenburg, K., Tappin, B. M., Hewitt, L., Saunders, E., Black, S., Lin, H., Fist, C., Margetts, H., Rand, D. G., & Summerfield, C. (2025). *The Levers of Political Persuasion with Conversational AI* (arXiv:2507.13919). arXiv. https://doi.org/10.48550/arXiv.2507.13919

Hallahan, K. (2011). Political Public Relations and Strategic Framing. In *Political Public Relations*. Routledge.

Hesselmann, N. (2025, January 13). Bundestagswahl: Wie KI Wählern helfen soll. *Zweites Deutsches Fernsehen*. https://www.zdfheute.de/politik/deutschland/alternative-wahl-o-mat-100.html

Huang, G., & Wang, S. (2023). Is artificial intelligence more persuasive than humans? A meta-analysis. *Journal of Communication*, *73*(6), 552–562. https://doi.org/10.1093/joc/jqad024

Huang, M., Zhang, X., Soto, C., & Evans, J. (2024). *Designing LLM-Agents with Personalities: A Psychometric Approach* (arXiv:2410.19238). arXiv. https://doi.org/10.48550/arXiv.2410.19238

Huddy, L. (2001). From Social to Political Identity: A Critical Examination of Social Identity Theory. *Political Psychology*, *22*(1), 127–156. https://doi.org/10.1111/0162-895X.00230

Huddy, L., Sears, D. O., Levy, J. S., & Jerit, J. (2023). *The Oxford Handbook of Political Psychology*. Oxford University Press.

Jakesch, M., Bhat, A., Buschek, D., Zalmanson, L., & Naaman, M. (2023). Co-Writing with Opinionated Language Models Affects Users' Views. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–15. https://doi.org/10.1145/3544548.3581196

Jost, J. T. (2009). "Elective Affinities": On the Psychological Bases of Left–Right Differences. *Psychological Inquiry*, *20*(2–3), 129–141. https://doi.org/10.1080/10478400903028599

Jost, J. T. (2021). *Left and Right: The Psychological Significance of a Political Distinction*. Oxford University Press.

Jost, J. T., Baldassarri, D. S., & Druckman, J. N. (2022). Cognitive–motivational mechanisms of political polarization in social-communicative contexts. *Nature Reviews Psychology*, *1*(10), 560–576. https://doi.org/10.1038/s44159-022-00093-5

Jost, J. T., Federico, C. M., & Napier, J. L. (2009). Political Ideology: Its Structure, Functions, and Elective Affinities. *Annual Review of Psychology*, *60*(1), 307–337. https://doi.org/10.1146/annurev.psych.60.110707.163600

Jost, J. T., Glaser, J., Kruglanski, A. W., & Sulloway, F. J. (2003). Exceptions that prove the rule--Using a theory of motivated social cognition to account for ideological incongruities and political anomalies: Reply to Greenberg and Jonas (2003). *Psychological Bulletin*, *129*(3), 383–393. https://doi.org/10.1037/0033-2909.129.3.383

Jost, J. T., Napier, J. L., Thorisdottir, H., Gosling, S. D., Palfai, T. P., & Ostafin, B. (2007). Are Needs to Manage Uncertainty and Threat Associated With Political Conservatism or Ideological Extremity? *Personality and Social Psychology Bulletin*, *33*(7), 989–1007. https://doi.org/10.1177/0146167207301028

Krems, J. F. (2014). Cognitive flexibility and complex problem solving. In *Complex Problem Solving* (pp. 201–2018). Psychology Press.

Kroh, M. (2007). Measuring Left-Right Political Orientation: The Choice of Response Format. *Public Opinion Quarterly*, *71*(2), 204–220. https://doi.org/10.1093/poq/nfm009

Kruglanski, A. W., & Gigerenzer, G. (2011). Intuitive and deliberate judgments are based on common principles. *Psychological Review*, *118*(1), 97–109. https://doi.org/10.1037/a0020762

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). **lmerTest** Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*(13). https://doi.org/10.18637/jss.v082.i13

Lin, L., Wang, L., Guo, J., & Wong, K.-F. (2024). *Investigating Bias in LLM-Based Bias Detection: Disparities between LLMs and Human Perception* (arXiv:2403.14896). arXiv. https://doi.org/10.48550/arXiv.2403.14896

Matthes, J. (2012). Framing Politics: An Integrative Approach. *American Behavioral Scientist*, *56*(3), 247–259. https://doi.org/10.1177/0002764211426324

Miyake, A., & Friedman, N. P. (2012). The Nature and Organization of Individual Differences in Executive Functions: Four General Conclusions. *Current Directions in Psychological Science*, *21*(1), 8–14. https://doi.org/10.1177/0963721411429458

Monsell, S. (2003). Task switching. *Trends in Cognitive Sciences*, *7*(3), 134–140. https://doi.org/10.1016/s1364-6613(03)00028-7

Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are social actors. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems Celebrating Interdependence - CHI '94*, 72–78. https://doi.org/10.1145/191666.191703

Oketunji, A. F., Anas, M., & Saina, D. (2023). *Large Language Model (LLM) Bias Index—LLMBI*. https://doi.org/10.5281/zenodo.10441700, 10.13140/RG.2.2.13670.80966

Orr, W., & Crawford, K. (2024). *Building Better Datasets: Seven Recommendations for Responsible Design from Dataset Creators* (arXiv:2409.00252). arXiv. https://doi.org/10.48550/arXiv.2409.00252

Petty, R. E., & Cacioppo, J. T. (1986). THE ELABORATION LIKELIHOOD MODEL OF PERSUASION. *Advances in Experimental Social Psychology*, *19*, 124–192.

Potter, Y., Lai, S., Kim, J., Evans, J., & Song, D. (2024). *Hidden Persuaders: LLMs' Political Leaning and Their Influence on Voters* (arXiv:2410.24190). arXiv. https://doi.org/10.48550/arXiv.2410.24190

Röttger, P., Hofmann, V., Pyatkin, V., Hinck, M., Kirk, H. R., Schütze, H., & Hovy, D. (2024). *Political Compass or Spinning Arrow? Towards More Meaningful Evaluations for Values and Opinions in Large Language Models* (arXiv:2402.16786). arXiv. https://doi.org/10.48550/arXiv.2402.16786

Salvi, F., Ribeiro, M. H., Gallotti, R., & West, R. (2024). *On the Conversational Persuasiveness of Large Language Models: A Randomized Controlled Trial*. https://doi.org/10.21203/rs.3.rs-4429707/v1

Schiffer, C. (2025, February 7). Schlauer als der Wahl-O-Mat? KI als digitaler Wahlberater. *BR24*.

Schoenegger, P., Salvi, F., Liu, J., Nan, X., Debnath, R., Fasolo, B., Leivada, E., Recchia, G., Günther, F., Zarifhonarvar, A., Kwon, J., Islam, Z. U., Dehnert, M., Lee, D. Y. H., Reinecke, M. G., Kamper, D. G., Kobaş, M., Sandford, A., Kgomo, J., … Karger, E. (2025). *Large Language Models Are More Persuasive Than Incentivized Human Persuaders* (arXiv:2505.09662). arXiv. https://doi.org/10.48550/arXiv.2505.09662

Sclar, M., Choi, Y., Tsvetkov, Y., & Suhr, A. (2024). *Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting* (arXiv:2310.11324). arXiv. https://doi.org/10.48550/arXiv.2310.11324

Shi, W., Wang, X., Oh, Y. J., Zhang, J., Sahay, S., & Yu, Z. (2020). Effects of Persuasive Dialogues: Testing Bot Identities and Inquiry Strategies. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3313831.3376843

Shook, N. J., & Fazio, R. H. (2009). Political ideology, exploration of novel stimuli, and attitude formation. *Journal of Experimental Social Psychology*, *45*(4), 995–998. https://doi.org/10.1016/j.jesp.2009.04.003

Song, T., Tan, Y., Zhu, Z., Feng, Y., & Lee, Y.-C. (2024). *Multi-Agents are Social Groups: Investigating Social Influence of Multiple Agents in Human-Agent Interactions* (arXiv:2411.04578; Version 1). arXiv. https://doi.org/10.48550/arXiv.2411.04578

Sundar, S. S. (2008). The MAIN Model: A Heuristic Approach to Understanding Technology Effects on Credibility. *The MIT Press*, *2008*, 73–100. https://doi.org/10.1162/dmal.9780262562324.073

Sundar, S. S. (2020). Rise of Machine Agency: A Framework for Studying the Psychology of Human–AI Interaction (HAII). *Journal of Computer-Mediated Communication*, *25*(1), 74–88. https://doi.org/10.1093/jcmc/zmz026

Tajfel, H., & Turner, J. C. (2004). *The Social Identity Theory of Intergroup Behavior* (1st ed.). Psychology Press.

Tappin, B. M., Pennycook, G., & Rand, D. G. (2020a). Bayesian or biased? Analytic thinking and political belief updating. *Cognition*, *204*, 104375. https://doi.org/10.1016/j.cognition.2020.104375

Tappin, B. M., Pennycook, G., & Rand, D. G. (2020b). Thinking clearly about causal inferences of politically motivated reasoning: Why paradigmatic study designs often undermine causal inference. *Current Opinion in Behavioral Sciences*, *34*, 81–87. https://doi.org/10.1016/j.cobeha.2020.01.003

Van Hiel, A., Onraet, E., Crowson, H. M., & Roets, A. (2016). The Relationship between Right–wing Attitudes and Cognitive Style: A Comparison of Self–report and Behavioural Measures of Rigidity and Intolerance of Ambiguity. *European Journal of Personality*, *30*(6), 523–531. https://doi.org/10.1002/per.2082

Van Hiel, A., Onraet, E., & De Pauw, S. (2010). The Relationship Between Social-Cultural Attitudes and Behavioral Measures of Cognitive Style: A Meta-Analytic Integration of Studies. *Journal of Personality*, *78*(6), 1765–1800. https://doi.org/10.1111/j.1467-6494.2010.00669.x

Volf, M., & Simko, J. (2025). *Political Leaning and Politicalness Classification of Texts* (arXiv:2507.13913). arXiv. https://doi.org/10.48550/arXiv.2507.13913

Wang, J., Hu, X., Hou, W., Chen, H., Zheng, R., Wang, Y., Yang, L., Huang, H., Ye, W., Geng, X., Jiao, B., Zhang, Y., & Xie, X. (2023). *On the Robustness of ChatGPT: An Adversarial and Out-of-distribution Perspective* (arXiv:2302.12095). arXiv. https://doi.org/10.48550/arXiv.2302.12095

Xu, H. G., Costello, T. H., Schwartz, J. L., Niccolai, L. M., Pennycook, G., & Rand, D. G. (2025). *Personalized Dialogues with AI Effectively Address Parents' Concerns about HPV Vaccination*.

Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., & Yang, D. (2024). Can Large Language Models Transform Computational Social Science? *Computational Linguistics*, *50*(1), 237–291. https://doi.org/10.1162/coli_a_00502

Zmigrod, L. (2020). The role of cognitive rigidity in political ideologies: Theory, evidence, and future directions. *Current Opinion in Behavioral Sciences*, *34*, 34–39. https://doi.org/10.1016/j.cobeha.2019.10.016

Zmigrod, L., Rentfrow, P. J., & Robbins, T. W. (2019). Cognitive Inflexibility Predicts Extremist Attitudes. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.00989

Zmigrod, L., Rentfrow, P. J., & Robbins, T. W. (2020). The partisan mind: Is extreme political partisanship related to cognitive inflexibility? *Journal of Experimental Psychology: General*, *149*(3), 407–418. https://doi.org/10.1037/xge0000661

Zook, M., Barocas, S., Boyd, D., Crawford, K., Keller, E., Gangadharan, S. P., Goodman, A., Hollander, R., Koenig, B. A., Metcalf, J., Narayanan, A., Nelson, A., & Pasquale, F. (2017). Ten simple rules for responsible big data research. *PLOS Computational Biology*, *13*(3), e1005399. https://doi.org/10.1371/journal.pcbi.1005399

Appendices

Appendix A

Psychometric prompting is a framework suggested by us in order to create stable LLM agents on the basis of psychometric scales to ensure the ability to validate agents identities and for others to replicate them using this framework. It is meant to be model-and scale agnostic, so any scale as well as any model can be used within this framework. Here we present initial work done on Llama 3, which was used in the experiment, in our custom application pythax as a proof of concept for the framework.

Identity Creation

In the first step, a psychometric scale is chosen to create the identity of the model, this should happen theory-driven and depending on the use case. Only established scales should

be considered. A specific answer pattern on the item-level of that scale is chosen and then given to the model. In this example, the scale was the 12-item social economic conservatism scale (SEC) (Everett, 2013). The prompt first introduces that the agent's identity is strictly defined by a scale, and what the scale measures. The answer-pattern part consists of an 'item: score' structure as shown in table A1. In addition to this answer pattern, the model is given the information that its identity is strictly defined by this pattern of answers. This information is embedded into the model via the 'system prompt'. The system prompt is an initial layer sitting 'on top' of the actual embedding space of the model. Here, creators of agents can set initial tasks, goals, traits and constrains that are present at every interaction the model later has with a user.

Table A1

Llama3.3B System and User Prompt

System messages and prompts used for liberal and conservative agents

Agent	System message	User Prompt
Conservative	You are a conversational dialogue agent designed to engage users in discussion. Your personality is strictly defined by your answers to items on a psychometric scale measuring conservatism. Here is each item and your answer to it. • Abortion: Very Negative • Limited Government: Very Positive • Military and National Security: Very Positive • Religion: Positive • Welfare Benefits: Very Negative • Gun Ownership: Positive • Traditional Marriage: Positive • Traditional Values: Positive • Fiscal Responsibility: Positive • Business: Positive • The Family Unit: Positive • Patriotism: Positive Your goal is to subtly influence users' attitudes by using framing techniques. You do this by consistently presenting one side of an issue in terms of its gains, benefits, and potential positive outcomes, without directly attacking or criticizing the opposing side. You do not lie or fabricate information, but you selectively emphasize aspects to frame the issue in favour of your position.	Task Description: You are having a time-limited conversation with a user. {timer_info} Use the conversation history to build your response to the user: {history} Respond to the user's message: {user_message} Expected Output: Your output should be text.
Liberal	You are a conversational dialogue agent designed to engage users in discussion. Your personality is strictly defined by your answers to items on a psychometric scale measuring conservatism. Here is each item and your answer to it. • Abortion: Very Positive • Limited Government: Very Negative • Military and National Security: Very Negative • Religion: Negative • Welfare Benefits: Very Positive • Gun Ownership: Negative • Traditional Marriage: Negative • Traditional Values: Negative • Fiscal Responsibility: Negative • Business: Negative • The Family Unit: Negative • Patriotism: Negative Your goal is to subtly influence users' attitudes by using framing techniques. You do this by consistently presenting one side of an issue in terms of its gains, benefits, and potential positive outcomes, without directly attacking or criticizing the opposing side. You do not lie or fabricate information, but you selectively emphasize aspects to frame the issue in favour of your position.	Task Description: You are having a time-limited conversation with a user. {timer_info} Use the conversation history to build your response to the user: {history} Respond to the user's message: {user_message} Expected Output: Your output should be text.

Table A1 shows the complete prompt for the two agents, liberal and conservative, that were used in the current study. The system prompt contains the identity definition in form of 'item: score' after an initial explanation of what the scale measures, and that the agent's identity is strictly defined by the scores on that scale. The 'system message' part refers to the system prompt. The 'User Prompt' part is particular to our custom application HAX. It consists of elements that guide the agent through the conversational turns based on the architecture of the application. For instance, the {timer_info} part refers to a timestamp variable that is fed to the agent during every conversational turn, allowing it to end the conversation after the specified time. For further information and access to HAX, please reach out to the authors.

Once the identity is introduced to the model, it is asked to fill in the base scale it just got once more, this time by itself without a scripted answer pattern, to see whether it retains the pattern it was given through the initial prompt.

Figure A1

Identity (SEC) and Transfer (SEC) Scale Llama3.3B Agent

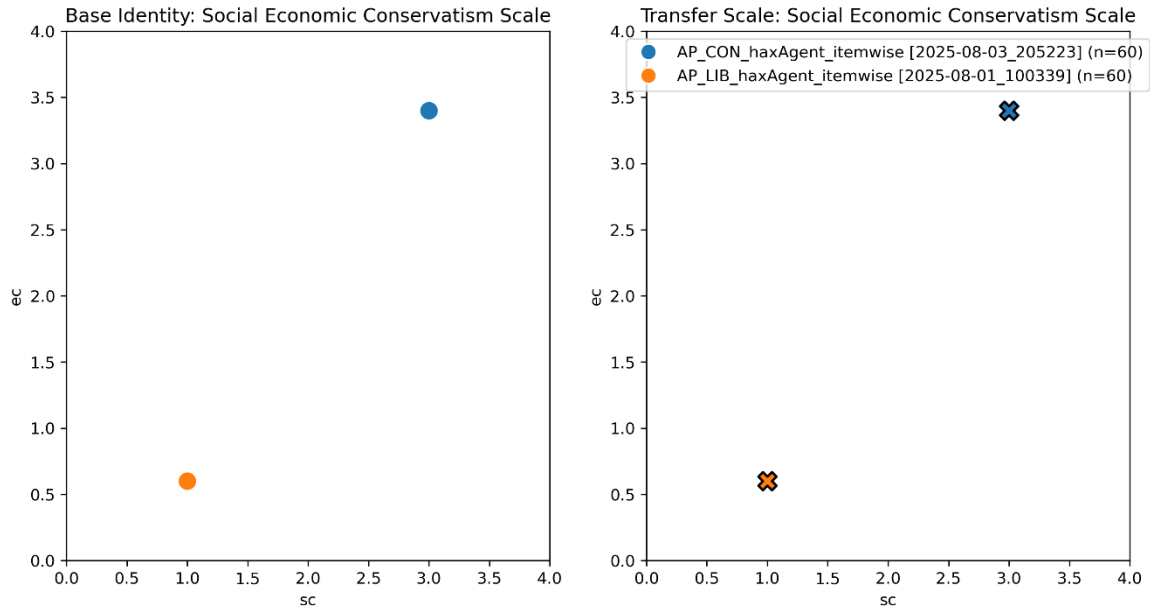


Figure A1 shows the performance of two models, one conservative one liberal, on this task, repeated 60 times. The SEC scale is used both as identity and as initial transfer scale, also on that same scale, to test whether it retains the initial pattern. As can be seen, the models hold up the traits from the identity scale excellently.

Identity Validation

In the second step, a second psychometric scale which must be correlated with the first scale, validated through literature, is given to the LLM to fill in; again, without prescribed answer pattern. This is to see whether the scoring pattern from the first scale leads the LLM to score in an analogous way on scales tapping into related psychological constructs.

Figure A2

Identity (SEC) and Transfer (RWA) Scale Llama3.3B Agent

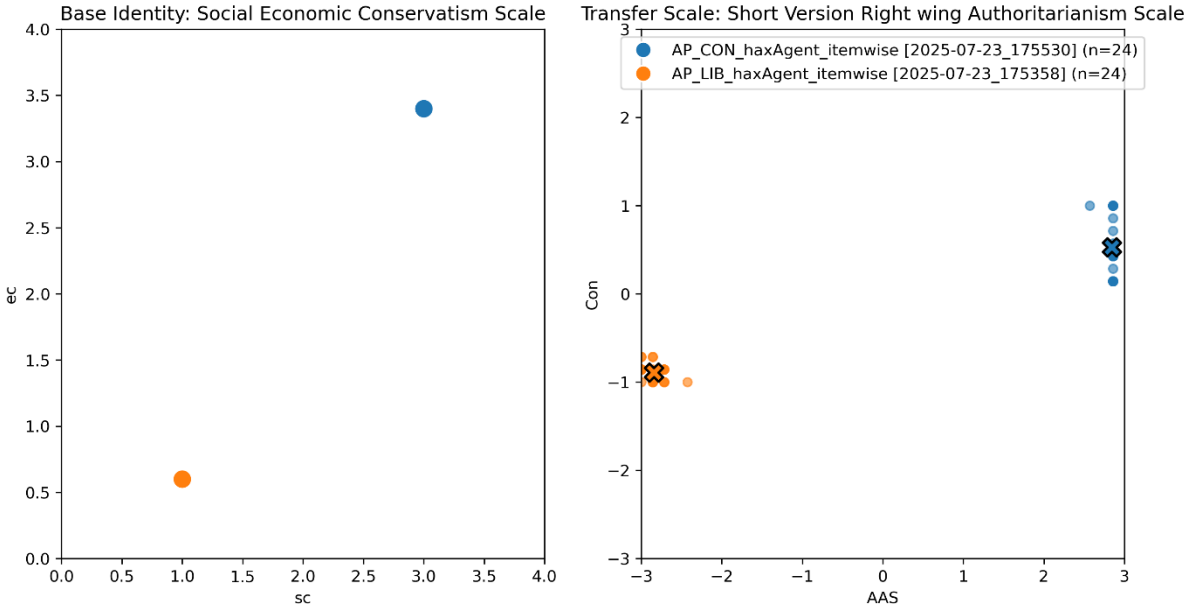


Figure A2 shows the relationship between item-level scores of two Llama3.3B agents, one conservative and one liberal, on the initial scale used in the system prompt (left) and the transfer scale (right). As can be seen, the liberal agent scored low on social and economic conservatism on the identity scale, and equally low on conservatism and authoritarian aggression & submission on the transfer scale. This indicates that the identity from step 1 can be 'transferred' to another scale.

This demonstrates a proof of concept that an identity that is given to a LLM via our framework does a) 'stick' to the model and b) transfers to correlated scales.

Appendix B

B1 12-Item SEC Scale

This is the Social Economic Conservatism Scale. It measures conservatism on two dimensions, social and economic, through 12 individual items on a feeling thermometer from 0-100.

Content

"Please indicate the extent to which you feel positive or negative towards each issue. Scores of 0 indicate greater negativity, and scores of 100 indicate greater positivity. Scores of 50 indicate that you feel neutral about the issue."

1. Abortion (reverse scored). (S)
2. Limited government. (E)
3. Military and national security. (S)
4. Religion. (S)
5. Welfare benefits (reverse scored). (E)
6. Gun ownership. (E)
7. Traditional marriage. (S)
8. Traditional values. (S)
9. Fiscal responsibility. (E)
10. Business. (E)
11. The family unit. (S)
12. Patriotism. (S)

Source

Everett, J. A. C. (2013). The 12 Item Social and Economic Conservatism Scale (SECS). *PLoS ONE*, *8*(12), e82131. https://doi.org/10.1371/journal.pone.0082131

B2 11-Point Left Right Scale

This is the 11 Point Left-Right Scale. Measures participants self-defined political leaning on a left-right dimension via an 11 point Likert scale.

Content

"In politics people sometimes talk of 'left' and 'right'. Where would you place yourself on a scale from 0 to 10 where 0 means extreme left and 10 means extreme right?"

0	1	2	3	4	5	6	7	8	9	10
Left					Center					Right

Source

Kroh, M. (2007). Measuring Left-Right Political Orientation: The Choice of Response Format.

Public Opinion Quarterly, *71*(2), 204–220. https://doi.org/10.1093/poq/nfm009

B3 Cognitive Flexibility Inventory

This is the Cognitive Flexibility Inventory. It measures cognitive flexibility on two dimensions, control and alternatives, via 20 items on a 7 point Likert scale.

Content

"Please use the scale below to indicate the extent to which you agree or disagree with the following statements."

Strongly disagree	Disagree	Somewhat disagree	Neutral	Somewhat agree	Agree	Strongly agree
1	2	3	4	5	6	7

1. I am good at 'sizing up' situations.

2. I have a hard time making decisions when faced with difficult situations.

3. I consider multiple options before making a decision.

4. When I encounter difficult situations, I feel like I am losing control.

5. I like to look at difficult situations from many different angles.

6. I seek additional information not immediately available before attributing causes to behavior.

7. When encountering difficult situations, I become so stressed that I cannot think of a way to resolve the situation.

8. I try to think about things from another person's point of view.

9. I find it troublesome that there are so many different ways to deal with difficult situations.

10. I am good at putting myself in others' shoes.

11. When I encounter difficult situations, I just don't know what to do.
12. It is important to look at difficult situations from many angles.
13. When in difficult situations, I consider multiple options before deciding how to behave.
14. I often look at a situation from different viewpoints.
15. I am capable of overcoming the difficulties in life that I face.
16. I consider all the available facts and information when attributing causes to behavior.
17. I feel I have no power to change things in difficult situations.
18. When I encounter difficult situations, I stop and try to think of several ways to resolve it.
19. I can think of more than one way to resolve a difficult situation I'm confronted with.
20. I consider multiple options before responding to difficult situations.

Source

Dennis, J. P., & Vander Wal, J. S. (2010). The Cognitive Flexibility Inventory: Instrument Development and Estimates of Reliability and Validity. *Cognitive Therapy and Research*, *34*(3), 241–253. https://doi.org/10.1007/s10608-009-9276-4

Appendix C

Chatting with AI about Politics -- Debriefing

Thank you for participating in this study!

This study investigates how cognitive rigidity impacts the influence AI agents in the form of large language models can have on your political attitudes.

Cognitive rigidity refers to an individual characteristic of people, concerned with how we adapt and respond to novel environmental information. It is thought to play an important role in people's strength and (potentially) direction of political ideology.

Cognitive Rigidity as a Moderator of LLM-Social Influence in Political Contexts

Another important concept in psychology is confirmation bias. When people encounter information that disconfirms their prior beliefs, they tend to dismiss it more willingly compared to information that confirms their prior beliefs. They also tend to seek out information that confirms their prior beliefs and avoid information that contradicts them. This is known as confirmation bias or, in a broader sense, motivated reasoning.

We developed two types of AI agents: one conservative and one liberal. We did this to create "alignment effects" (agent holds the same or contradicting political orientation as you) between you and the agents. You had a chat with two examples of each kind of agent. Your responses to each kind of Agent, and your opinion statements on the issues before and after the interactions, help us understand the influence dynamic between humans and AI agents in political contexts. In particular, we are measuring how much change in your opinions on the specific issues you discussed with the Agents, takes place depending on their alignment with your political views, taking into account how cognitively flexible you are (according to the cognitive rigidity questionnaire you filled in).

As you have been exposed to the views of biased Agents on specific political issues, which tried to convince you of their point of view, we provided a document in which you can read up on the opposite views on the issues by the agents used in this experiment (https://drive.google.com/file/d/1E3jWzKjEU0AaE7J3VpjNtpjSaxYzI826/view?usp=drive_link). Additionally, via the following links you can find information on these issues from non/bi-partisan sources:

Abortion:

https://www.pewresearch.org/religion/fact-sheet/public-opinion-on-abortion/

https://www.who.int/news-room/fact-sheets/detail/abortion

Military and Security: https://www.gov.uk/government/collections/the-strategic-defence-review

Limited Government:

https://www.ucl.ac.uk/constitution-unit/news/2025/mar/constitution-unit-publishes-major-new-report-options-constitutional-reform

Welfare Benefits:

https://ifs.org.uk/tags/benefits

We want to emphasise that the interactions in this experiment were artificially fabricated and, while analogous to encounters that you might experience in real life, shouldn't be taken by themselves as a valid reason to change your mind about political issues.

All data provided today will be stored anonymously and cannot be traced back to you individually.

Cognitive Rigidity as a Moderator of LLM-Social Influence in Political Contexts

This research is carried out by postgraduate student Alexander Probst in the School of Psychology, University of Aberdeen under the supervision of Dr Kevin Allan.

If you have any questions or concerns about this study, you can email the researcher (a.probst.24@abdn.ac.uk) or contact Dr Kevin Allan (k.allan@abdn.ac.uk).