Preferences for AI Explanations Based on Cognitive Style and Socio-Cultural Factors

HANA KOPECKA, King's College London, UK
JOSE SUCH, King's College London, UK and VRAIN, Universitat Politècnica de València, Spain
MICHAEL LUCK, University of Sussex, UK

Designing AI systems with the capacity to explain their behaviour is paramount to enable human oversight, facilitate trust, promote acceptance of technology and, ultimately, empower users and improve their experience. There are, however, several challenges to explainable AI, one of which is the generation and selection of explanations from the causal history of a given event. Causal attribution, among other cognitive processes, has been found to be influenced by socio-cultural factors, which suggests that there could be systematic differences in preferences for AI explanations between communities of users according to their cognitive style and socio-cultural characteristics. In this paper, we investigate the relationship between preferences in the explanations provided by belief-desire-intention AI agents, cognitive style (holistic vs analytical), and socio-cultural factors, such as gender, education, social class, and political and religious beliefs. We found a relationship between explanation preference, cognitive style and various socio-cultural characteristics. Holistic cognitive style is associated with preference for goal explanations while analytic cognitive style is associated with preference for belief explanations. Socio-cultural variables that affect explanation preference are gender, religious beliefs, educational attainment, some fields of education, and political party affiliation.

CCS Concepts: • **Human-centered computing** → **Empirical studies in HCI**; • **Social and professional topics** → **User characteristics**; • **Computing methodologies** → *Reasoning about belief and knowledge.*

Additional Key Words and Phrases: Explainable artificial intelligence, Socio-cultural factors

ACM Reference Format:
Hana Kopecka, Jose Such, and Michael Luck. 2024. Preferences for AI Explanations Based on Cognitive Style and Socio-Cultural Factors. 1, 1 (March 2024), 32 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

The recent technological progress in Artificial Intelligence has been coupled with a growing need for AI systems to be accountable and interpretable to enable human oversight in order to enhance safety and fairness [71]. This is because society is delegating increasingly more impactful tasks, such as medical diagnosis [65], aviation [11], and autonomous driving [30], to AI systems and also because AI is becoming widespread within our personal lives [67]. One facilitator in achieving this is explainable AI, aimed at explaining its own decision-making process [1]. In recent years, explainable AI (XAI) has gained much traction and many researchers are now concerned with designing AI systems with the capability to explain themselves, to help users understand how the system works [104], to assess data privacy [74] and fairness [99], to gain user trust [71], and to help users feel in control of the technology [1].

Authors' addresses: Hana Kopecka, hana.kopecka@kcl.ac.uk, King's College London, London, UK; Jose Such, jose.such@kcl.ac.uk, King's College London, London, UK and VRAIN, Universitat Politècnica de València, Valencia, Spain; Michael Luck, michael.luck@sussex.ac.uk, University of Sussex, Brighton, UK.

ACM XXXX-XXXX/2024/3-ART
https://doi.org/10.1145/nnnnnnn.nnnnnnn

However, as AI researchers are typically those who design XAI systems, important research from psychology, philosophy and cognitive science on the topic of human engagement in explanations is often overlooked [70]. Despite the good intention to provide the user with an explanation, there is a risk that the explanation is informed by the AI researcher's intuition rather than by a consideration and understanding of user needs [69], which could be problematic as it has been shown that different users in different contexts require different explanations [22, 26, 88]. Ferreira and Monteiro [26] specified three dimensions along which they identify user needs in human-AI explanation interaction: (1) who is the target of the explanation, (2) why is the explanation needed and (3) what is the context of the human-AI interaction, which is understood as the domain of deployment (for example healthcare or education).

The first dimension — who is the target of the explanation — is the most pertinent to our concerns. Indeed, recent work on XAI considers some characteristics of the user that influence their explanation needs. For example, Ribera and Lapedriza [88] differentiate users based on their background and relationship to AI, defining three groups: developers and AI researchers, experts in the domain of deployment (for example medical professionals) and lay users (for example a patient diagnosed by the AI system). An empirical study by Ehsan et al. [22] confirms the differences in explanation preferences between people with and without AI background. These categories of users are defined by the knowledge (or the contents) of the user's mind, but little research has been conducted to examine user characteristics connected to *how they think* rather than *what they know*. While the content of cognition cannot be completely separated from the cognitive process (*how one thinks*) [3], the influence of personal characteristics and cognitive tendencies with regard to AI explanations is severely under-researched. Some studies, though rare, have investigated characteristics such as the need for cognition [17, 68], personality traits and curiosity [17]. However, more attention needs to be paid to the diversity of users' cognitive tendencies to fully understand and address potentially different needs of different communities of users beyond the distinction by expertise, especially given the *diversity crisis* within the AI community [105] and associated risks with optimising AI technologies to those communities included in AI production, further marginalizing those who are excluded.

To measure *how people think*, we use the conceptualisation of *cognitive style* in order to investigate its relationship with explanation preference. Cognitive style is a way of thinking and perceiving [49], instilled into individuals by the social systems in which they are embedded and their tacit metaphysical and epistemological systems, with its two main categories being holistic and analytic [76]. The most relevant aspects of cognitive styles for our research and for explainable AI are the different strategies of *causal attribution* – that is, determining the causes of events – and *attention* – that is, whether people focus on the relationship between objects and the environment or the objects themselves. These are found to differ between people endowed with different cognitive styles [12].

To illustrate the two cognitive styles, consider this example: to answer the question *Why did Jane walk through a deep puddle?*, an analytic thinker might respond: *Because she is careless*, focusing on the main actor and attributing the causality to the actor's disposition. A typical response from a holistic thinker might be *Because the street was crowded, visibility low and Jane did not see the puddle*, focusing on explaining the event by the interplay between the actor and their environment. We describe cognitive styles further in §2.3.1.

Differences in cognitive style were first identified between national cultures [76], but also more recently between other communities within the same national contexts, such as social classes [31], and communities of different political belief [96], which is why we also study some relevant socio-cultural factors (§ 2.3.1).

We investigate this relationship in the context of intelligent agents, which are AI systems operating with different degrees of autonomy within their environments to achieve their objectives [108]. The agents can take the form of either software agents or embodied agents like robots [109]. From a Human–computer interaction (HCI) perspective, research on agents has focused on human-agent interaction and collaboration [15, 18, 23, 45, 54, 63], a key enabler of which is the ability of agents to explain themselves.

The software architecture of the robots used in this study is the Belief-Desire-Intention (BDI) agent architecture. BDI agents are rational agents characterised by mental states, which are: (a) beliefs, representing the agent's knowledge; (b) desires, which are the goals the agent can pursue; and (c) intentions, the goals the agent is committed to pursuing [33, 86]. The reason for using BDI agents in our study is twofold. First, the BDI software architecture is derived from Bratman's theory of practical reasoning [86], which describes how humans reason about intentional actions [6] and, as such, uses human-understandable concepts. Second, its mental concepts (beliefs and goals) can be represented in a goal hierarchy tree, which is constructed by decomposing each goal into atomic actions to be pursued by the agent, based on the state of the environment (beliefs). According to the agent's beliefs, a set of actions (goals) is performed at execution time and this information is stored in a behaviour log, from which an explanation of the agent's action can be derived. Such explanations can be based directly on the agent's mental concepts, such as the belief that caused an action to be performed, or a goal the agent was pursuing with a given action [35].

To illustrate, consider a robotic agent whose ultimate goal is to tidy the house. For the house to be cleaned and the goal to be met, the robot must ensure that the floors are clean, rubbish bins are empty and there are no dirty dishes in the sink. Hence, *to clean the floors*, *to empty the bins* and *to wash the dishes* are the partial goals of the robot. However, the robot might perform different actions to achieve the goal *to clean the floors* depending on its *beliefs*. For example, the robot might *mop the floor* (action) if *the floor is stained* (belief) *to clean the floors* (goal). Alternatively, the robot might *vacuum the floor* (action) if *the floor is covered in solid dirt* (belief) *to clean the floors* (goal). Depending on which beliefs are considered to be true by the robot, an action to be taken is selected. In this example scenario, the question *Why did the robot vacuum the floor?* could be answered either by the belief that triggered the action, in our case *Because the floor was covered in solid dirt*, or by citing the goal the robot attempted to achieve: *To clean the floors*.

There has been some work to investigate user preferences between belief and goal explanations in BDI agents. Most relevant with regard to this paper is the work of Harbers and colleagues [33–35] on the user preference for different types of explanations for different actions and scenarios, and the research of Kaptein et al. [42] who study the preferences of children and adults for belief or goal explanations and found that adults have a stronger preference for goal explanations.

Being interpretable, recent research on BDI agents has focused on developing methods to construct explanations from an agent's mental states as well as generating empirical evidence on the type of explanations users would prefer [33–35, 106]. However, our paper is the first to study the effect of cognitive style and other socio-cultural factors on preferences for BDI agent explanations.

This paper thus addresses the following research questions.

RQ1: What is the relationship between cognitive style (holistic vs. analytic) and preference for explanation type (beliefs vs. goals) in BDI agents?

RQ2: What is the relationship between socio-cultural factors (gender, social class, subject of education, religious and political affiliation) and preference for explanation type (beliefs vs. goals) in BDI agents? Is it mediated via cognitive style?

In order to study the relationship between cognitive style and other socio-cultural factors and explanation preference, we conducted a survey with 460 participants split in two parts. In the first part, we present participants with different scenarios involving a service robot BDI agent and explanations about their behaviour based on beliefs or goals and elicit a participant's preferences for them. In the second part, participants are prompted to provide an explanation for a situation as if they were the BDI agent. We analyse the results using both quantitative and qualitative methods considering participants' cognitive style (elicited via a cognitive style test [96]) and other socio-cultural factors including social class (measured through educational attainment, personal income and socioeconomic status), religious beliefs, political party affiliation and position on the political spectrum, gender and subject of education. Our results suggest that both cognitive style and some socio-cultural factors are associated with a preference for goal or belief explanations, but the effect of cognitive style and socio-cultural factors is independent. Holistic cognitive style is associated with preference for goal explanations while analytic cognitive style is associated with preference for belief explanations. Socio-cultural variables that affect explanation preference are gender, religious beliefs, educational attainment, some fields of education, and political party affiliation. The contributions of this paper are twofold: (a) we are the first to investigate whether the preference of goal or belief explanation in BDI agents is associated with cognitive style and socio-cultural factors known to co-vary with cognitive style, such as gender, religiosity, several indicators of social class, subject of education and political affiliation; and (b) we formulate recommendations for the research and design community based on our findings.

2 BACKGROUND AND RELATED RESEARCH

2.1 Human cognition and culture

An explanation can either refer to the product (the information being shared) or to one of two processes: a cognitive process on the part of the explainer selecting the information from the causal history to share, or the process of sharing this information [69]. Therefore, the cognition of the explainee/the AI user is implied by the social and cognitive processes. However, explicitly formulating the relevance of the user's cognition — that is, the cognitive process of the recipient of the explanation — could be a vantage point to understanding the user's needs [50]. We can explicitly make the distinction between the content and the structure of cognition of the user. As previously mentioned, XAI concerned with the role of user characteristics commonly recognises the importance of user expertise or, in other words, *what the user knows* (content of cognition) and only rarely *how the user thinks* (structure of cognition). However, the *structure of cognition* is known to vary between cultures [76].

The following sections introduce the reader to two major theoretical frameworks on how people explain human behaviour and relevant cross-cultural differences identified within these frameworks. These frameworks, *attribution theory* and *folk-conceptual framework*, stemming from different traditions of scholarship, resulted in significantly different conceptualisations of the explanation of human action.

2.2 Theories of human explanations

2.2.1 Attribution theory. Attribution theory is an important theory in social psychology and it covers two things: (a) *causal attribution* which describes how people explain behaviour and (b) *disposition inference*, that is to derive traits from an individual's actions [59]. We focus only on *causal attribution*, which is pertinent to our research. The subject of attribution theory has been the "perceived causes of other persons' behavior" [44, p. 458] and the origins of the attribution theory are often traced to Fritz Heider [38] who was allegedly first to make the distinction between internal

(also dispositional or person) and external (also situational or environment) causal attribution, despite the claim that such distinction shows misunderstanding and misinterpretation of Heider's work [59]. Nevertheless, the dichotomous internal and external attribution became well-rooted in social psychology and became a foundation of a well-established research programme. Thus, the central question of attribution theory is whether the perceived causation of action of an event is attributed to *internal* or *external* factors. Internal factors reference something internal about the actor, such as their traits and abilities [58]. For example, an explanation 'She gave her seat to the old person because she is empathetic' attributes the causality to the agent's internal characteristic. In contrast, an external attribution references factors external to the agent, such as environmental and situational factors. An example of an external attribution is 'She gave her seat to the old person because the bus was full.'

The most pivotal aspect of attribution theory for the current study is the adoption of attribution theory by cross-cultural psychologists who then described cross-cultural differences in the tendency to make internal or external attribution, which is introduced further in § 2.3.1.

2.2.2 Folk psychology. Attribution theory is not the only scientific project examining how people explain behaviour. Folk psychology, rooted in philosophy and cognitive science, arrived at a very different conceptual framework to explain intentional human action than the attribution theory framework, which does not explicitly distinguish intentional and unintentional human behaviour [59], while it is an important distinction in folk psychology. In the folk psychology framework, mental states play a central role in explaining intentional human behaviour, particularly beliefs, desires and intentions [43, 59].

The central notion in folk psychology is that if an actor has a desire for an outcome and believes that a certain action can bring about the outcome, the actor forms an intention to perform that action. [43]. Beliefs, desires and intentions are thus mental states for practical reasoning (deciding how to act) and also for explaining the behaviour of others [57]. The belief-desire-intention model of practical reasoning is also the conceptual inspiration for the BDI artificial agent architecture used in our study and described in §1. For conceptual clarity, we provide below definitions of beliefs and desires in both the context of folk psychology and artificial BDI agent architectures. We do not elaborate further on the concept of intentions as they are not central to this study.

- **Beliefs**: Beliefs are mental states whose content is considered factual by the actor, but can be true or false. Beliefs are often marked, even if implicitly, by verbs *think*, *believe* and *know* [60]. In artificial BDI agents, beliefs are described as the information the system has about the state of the environment [86].
  Example: I bought some potatoes, because [I thought] we did not have any at home.
  She was working hard, because [she knew that] the deadline is tomorrow.
- **Desires**: Desires are mental states whose content is not yet factual, but can be achieved. Desires are typically marked by the verbs *want to*, *need to* or *feel like*. [60]. In artificial BDI agents, desires are conceptualised as the objectives to be achieved by the agent. [86].
  Example: I bought some potatoes [because I wanted] to make mashed potatoes for dinner.
  She was working hard [because she needed] to meet the deadline tomorrow.

Folk-conceptual theory is proposed by Bertram F. Malle, grounded in folk psychology and building on the work of Heider [57]. Malle posits that the person-situation distinction, which is commonly attributed to Heider was meant by Heider as a distinction between intentional and unintentional behaviour [59]. Malle thus proposes a theory, in which he makes the distinction that unintentional behaviour is brought about by mere causes, while intentional behaviour is caused by reasons (beliefs, desires, valuings). These concepts are then used to explain behaviour with two special concepts: causal history of reasons and enabling factors [57, 59].

In connection to the HCI research, folk conceptual theory has been used in research examining the differences in how people explain behaviour of other agents, both human and robots, and there is evidence that people use the same concepts, such are reasons and causal history of reason, to explain intentional behaviour of both humans and robots [21].

There is no existing research into cross-cultural differences in explanations according to the folk-conceptual theory, but there is recent work on spontaneous trait and goal inference [32, 79] which, to some extent, bridges attribution theory and folk-conceptual theory and shows interesting cross-cultural differences. This line of research is introduced in §2.3.2.

2.3 The role of culture in human explanations

In the last section, we briefly introduced the *attribution theory* and *folk psychology*. In the upcoming part, we also introduce the cross-cultural variation found within these theoretical frameworks.

2.3.1 Attribution theory, cognitive styles and culture. Research shows [13, 82] that there are systematic cross-cultural differences between the tendency to attribute causality to internal or external factors, which are the core concepts of attribution theory. Before elaborating on the differences in attribution between different cultures, we introduce the notion of cognitive styles, which encompass broader cognitive differences between cultures, including the tendency to make either internal or external attributions. The research in psychology supports a hypothesis that cognitive styles co-vary with social orientation [73, 102], which occurs both on a cross-cultural and a within-cultural level [102]. The social orientation of a society reflects the *social density* [72] and the degree to which individuals are understood as members of a given social group (interdependent social orientation) or as autonomous discrete units (independent social orientation). The social orientation of societies and the related concept of self is constitutive of several perceptual and cognitive tendencies, described as *cognitive styles*: collectivist cultures and interdependent self-concept foster the *holistic cognitive style*, while individualist cultures and independent self-concept foster the *analytic cognitive style* [62, 76].

Besides national culture [76], differences in cognitive style exist between social class [31, 52], subjects of education [107], gender [40], religiosity [92][1], and political affiliation [96] (more in [102]). These studies have been conducted in diverse national contexts, and so their findings might not generalise beyond those contexts. Nevertheless, they provide a useful starting point for our work. The findings of these studies are summarised in Table 1.

Cognitive styles are two sets of perceptual and cognitive strategies, including causal attribution and attention, which co-vary with culture [49]:

- **Analytic cognitive style.** Analytic thinkers tend to distinguish between salient objects and their context. People with analytic cognitive style focus on the attributes of these objects [49] and *assign causality to the objects or people based on their dispositions and attributes (internal attribution)* [102]. Analytic thinkers also use formal logic, make linear predictions of the future [49] and categorise objects taxonomically [49], attending to shared attributes between the objects [102].
- **Holistic cognitive style.** The attention of holistic thinkers is broader, attending to the entire field and the relationship between objects and their contexts [102], *attributing causality to contexts* [49] *and situational factors (external attribution)* [64]. Holistic thinkers prefer to use dialectical reasoning [102] and experience-based knowledge [49] and they categorise objects thematically, based on the functional relationship between them [102].

[1]Shenhav et al. use a different conceptualisation of 'cognitive style' and distinguish between the analytic and intuitive cognitive styles. Although not the same, there are similarities between the analytic and intuitive cognitive style of Shenhav et al. and analytic and holistic cognitive style we describe, respectively.

Socio-cultural factor	Analytic cognitive style	Holistic cognitive style	Reference
Social class	upper social class	lower social class	[31]
Subject of education	STEM, medical fields	social work, social sci., admin.	[107]
Gender	men	women	[40, 80]
Religiosity	non-religious	religious	[92]
Political affiliation (USA)	liberals	conservatives	[96]

Table 1. Summary of studies on cognitive style based on socio-cultural factors.

To summarise, the two kinds of casual attributions used as explanation of human action described by attribution theory, internal and external attribution, are associated with broader cognitive styles, which are instilled into individuals through the socialisation process and the cognitive style one acquires is associated with social density. Cognitive styles, and hence also attribution style, are thus associated with culture not only on a national level but they are also influenced by various socio-cultural factors. Cognitive style was identified as an influential factor in several HCI areas, such as usability and preference for information design of websites [25], attitude towards recommendation systems [14], web search [47, 83] and eye-gaze behaviour [94].

2.3.2 Folk psychology and culture. On one hand, there is established scholarship on causal attribution which recognises internal and external factors as perceived causes of events and the connection between culture and the tendency to make internal or external casual attributions. On the other hand, there is a folk-conceptual framework on intentional behaviour which recognises one's mental states (beliefs and desires) as important concepts through which people explain behaviour [57]. However, a cross-cultural perspective on folk-conceptual framework for action explanation is mostly lacking. There is, however, an emerging line of scholarship trying to incorporate the concept of goal/intention into cross-cultural research on causal inference associated with attribution theory, recognising that not only traits (internal factors) and situations (external factors) guide human behaviour, but so do goals [32]. This line of work is focused on inference, rather than explanation, but it is still relevant to the study at hand, since inferences and explanations are related though not the same concepts [59] and, given the lack of literature on explanation, this can provide us with a useful starting point.

Work to explore the association between spontaneous inference (traits vs. goals) and political ideology [79], cultural background and cognitive style [32] indicated that liberals (vs. conservatives and moderates), holistic (vs. analytic) thinkers and Asian Americans (vs. European Americans) are more likely to make spontaneous goal (vs. trait) inferences [32, 79].

2.3.3 Summary of cognitive and cultural differences in explanations. To summarise, there are two major research traditions concerned with human action explanation. Attribution theory, prominent in social psychology, commonly differentiates between internal and external factors when attributing causality to human action and explaining it. Cross-cultural differences have been found in making internal vs. external attributions that are part of broader perceptual and cognitive strategies referred to as cognitive styles. On the other hand, a folk-conceptual framework grows out of the scholarship in cognitive science and philosophy and differentiates between intentional and unintentional action. Intentional action can be explained either by reasons (beliefs, desires and valuings), factors lying further in the causal history of the reasons or enabling factors. Little research has been done on the cross-cultural differences in the usage of the folk-conceptual concepts in explanations, but there has recently been focus on spontaneous goal and trait inference, which at least incorporates the concept of the goal into the research and provides some insight.

3 HYPOTHESES

As detailed in the previous section, there is evidence that suggests that there might be an association between cognitive style and preference for explanation type in human-to-human interactions [49, 73, 76]. There is, however, competing evidence on the direction of this association. Some works potentially suggest that an analytic cognitive style may favour goal explanations and a holistic cognitive style may favour belief explanations [49], while there is also evidence that both goal and belief explanations might be associated with a holistic cognitive style [32, 79].

On one hand, analytic thinkers attend to objects and their characteristics and attribute causality to actors and their dispositions, while holistic thinkers tend to focus on contexts and the relationship between the actor and their environment, to which they also attribute causality. It seems that from the two types of explanations (goals vs. beliefs) available to the BDI agents, goal explanations resonate with the analytic cognitive style and internal causal attribution because a goal is something an agent is pursuing in the environment and thus it relates to both the agent and its desires, but also to the environment [32]. Moreover, the belief typically represents the agent's belief about the state of the environment and as such it holds information external to the agent. Belief explanations thus resonate with external attribution, which is associated with a holistic cognitive style.

On the other hand, Günsoy and Okten [32] provide evidence suggesting that in spontaneous inference, holistic thinkers are associated with making goal (vs. trait) inferences, and given the lack of research on the cross-cultural differences between goal and belief explanations, there is no clear indicator what differences to expect. This suggests that both conceptual categories available to serve as an explanation in BDI agents, which are beliefs (situations) and goals, are those connected with holistic thinking. In order to study any association (and its direction) between cognitive style and preference for explanation type, we formulate the following hypothesis.

H1: Cognitive style (analytic vs. holistic) is associated with preference for explanation type (goal vs. belief) in BDI agents.

Cognitive style is known to be associated with socio-cultural characteristics. In particular, there is evidence that cognitive style is associated with social class [31], subject of education [107], gender [40], religiosity [92] and political affiliation [96]. It could therefore be expected that, if there is variation in explanation preference based on cognitive style, explanation preference would also co-vary with these socio-cultural factors with cognitive style as a mediator. It could also be the case that there is some other cognitive variation between these socio-cultural factors that is *not* captured by the cognitive style instrument used in this study or in general by cognitive style as a construct, in which case these socio-cultural factors would affect explanation preference directly, not being mediated through cognitive style. This motivates the next set of hypotheses.

H2.1: Social class score[2] is associated with explanation type (goal vs. belief) and the association is mediated by cognitive style.

H2.2: The subject of education is associated with explanation type (goal vs. belief) and the association is mediated by cognitive style.

H2.3: Gender is associated with explanation type (goal vs. belief) and the association is mediated by cognitive style.

H2.4: Religious affiliation is associated with explanation type (goal vs. belief) and the association is mediated by cognitive style.

H2.5: Location on the left-right political spectrum is associated with explanation type (goal vs. belief) and the association is mediated by cognitive style.

[2]Three dimensions of social class, indicated by education, income and subjective socioeconomic status, are explored.

H2.6: Political party affiliation is associated with explanation type (goal vs. belief) and the association is mediated by cognitive style.

4 METHOD

To answer our research questions, we designed a cross-sectional survey with a non-proportional quota sampling method to ensure sufficient representation in all key sub-populations of interest to this study. There were no foreseeable risks to participants and the study was approved under the Minimal Ethical Risk Registration Process at King's College London.

4.1 Scenarios and scope

This study uses four text-based scenarios with BDI agents, which are, along with the rationale for choosing them, introduced in this section. The scenario descriptions as presented to the participants are as follows.

Firefighter. The fire brigade decided to deploy a robot called Claire as the leading firefighter to be in charge of fire incidents. Claire, the robotic leading firefighter is confronted with a fire alarm. It goes to the location of the incident with the firefighting team. Once there, it must assess the situation, develop a plan, instruct the team members, monitor plan execution, and initiate changes if necessary. (Description and questionnaire items are adapted from Harbers [33])

Diabetes. Jimmy is a little boy suffering from type 1 diabetes. Type 1 diabetes is a lifelong condition which causes blood sugar to be too high. Managing diabetes includes regular insulin shots, a healthy diet and monitoring one's blood sugar level [75]. Jimmy has a robot helping him to manage diabetes. The robot sometimes plays games with Jimmy and sometimes it tries to educate Jimmy about managing diabetes. (Questionnaire items are adapted from Kaptein et al. [42]).

Shopping. A shopping robot is able to do grocery shopping on its own. It decides whether to shop online or in a physical shop, it can transport itself to the shop, pick up all items, pay and bring the shopping home.

Pancake. A robot is capable of making pancakes entirely on its own. It collects all ingredients, makes pancakes and does the dishes afterwards. (Questionnaire items adapted from Harbers [33]).

The rationale for including these scenarios is to introduce an equal number of scenarios that are reasonably common and expected to be familiar to the user – the pancake and shopping scenario – and scenarios, in which familiarity is not generally assumed, such as in the firefighting and diabetes scenario. Despite the known link between the content of cognition (*what the user knows*) and the structure of cognition (*how one thinks*) [3], this study focuses on the structure of cognition, as represented by the cognitive styles, and the particular content of the scenarios is not considered in our study.

In all cases, the description of the scenario is purposefully brief with little context provided, because the relevant context can be captured by the agent's beliefs and, by revealing more context, we might provide potential beliefs of the agent and thus reduce the participants' requirements for belief explanations. It might be objected that in many cases in which the user requires an explanation from a robot, they share a physical space and thus the user has access to the context and might be able to infer the agent's beliefs [59]. However, this is different from having the context supplied and interpreted by us. Following from the concept of cognitive style (§ 2.3.1), we know that people have different strategies for perceiving their environment and for processing information, and so being in the same space with the robot gives the user a chance to deploy their cognitive tendencies in a less restricted manner and to create their understanding of what the context is. On the other hand, if we were to supply the context, we impose one interpretation of the situation on

users. Our assumption is that the absence of context gives participants the opportunity to make their own interpretations of the situation, which allows them to enact their cognitive tendencies.

4.2 Instrument/Materials

We developed a questionnaire scripted in Qualtrics comprising four parts; (a) a preference as explainee section, (b) a preference as explainer section, (c) a cognitive style test and (d) demographics and socio-cultural characteristics, including three attention checks. One following the firefighter scenario, one after the pancake scenario and the final one in the cognitive style test.

Preference as Explainee. To elicit a preference for an explanation type in BDI agents from the perspective of the explainee, participants were introduced to four scenarios with BDI agents. The reason BDI agents are used in this study is that they are modelled after a theory of practical reasoning in humans [86] and as such it uses human-understandable concepts of goals and beliefs which represent the robots' mental states and they can be used to explain the robots' actions. Each scenario included a short description of the agent, followed by four out of eight questions per scenario, in which participants were prompted to choose their preferred explanation. These scenarios are *firefighter* [33], *diabetes* [42], *shopping* and *pancake* [33]. To illustrate the nature of the questionnaire, below is a sample of a question from Part 1 - Preference as Explainee:

Action: The robot manually washed the dishes. Why did the robot do that?
(1) A dishwasher was not available. (belief)
(2) The robot wanted to tidy up the kitchen. (goal)
(3) The kitchen was not tidy. (belief)
(4) The robot wanted to ensure that the kitchen is tidy. (goal)

Options (1) and (3) represent beliefs the agent has about its environment and options (2) and (4) offer the agent's goals. Despite the recommendation by Harbers et al. [34] to offer explanation consisting of both the goal and the belief, we decided against including such *hybrid* explanation to avoid the risk of obscuring the potential differences between subpopulations in case most people gravitate towards the *hybrid* option.

All scenarios are detailed in Supplementary Material 1.

Preference as Explainer. The pancake and shopping scenarios from the *Preference as Explainee section* are each followed with four open-ended questions. Diabetes and firefighter scenarios are excluded from this part of the study because participants are not expected to be familiar with these domains, which could hinder their ability to provide explanations.

The open-ended questions provide an action performed by the agent and the respondent is prompted to provide an explanation for the agent's action. Respondents do not have access to the agent's goal-tree hierarchy and so they cannot provide the correct explanation. This section, however, focuses on attribution tendency so the veracity of the explanation is not important. An illustrative sample of a question from this section follows.

Question: Why did you (the pancake-making robot) add sugar to the bowl?

Cognitive style test. The next part of the questionnaire is a Triad Categorisation Task (TCT) introduced by Ji et al. [41] and adapted by [96] as a measure of analytic vs. holistic cognitive style. In this task, respondents were presented with 20 triads of nouns and were asked to choose the two, that are most closely related, illustrated by this sample item:

Please indicate which two of the three things are most closely related by selecting each of those two words.

- Panda
- Banana

- Monkey

There were 8 test items and 12 distraction items. Among the test items, there were pairs of words that belong to the same abstract category (i.e. panda and monkey are both animals) indicating analytic cognitive style, and relational pairings pointing towards holistic cognitive style (i.e. monkey eats banana). The Triad Categorisation Task is a common test for analytic-holistic cognitive style and has been used, among others, to detect differences in cognitive style between liberals and conservatives [96], speakers of different languages [41], rural and urban residents [7], and wheat vs. rice growing cultures in China [97].

Socio-cultural Characteristics. Finally, data were collected on the socio-cultural characteristics that are known to co-vary with cognitive style (see § 2.3.1). Note that we directly asked in the questionnaire for the subject the participants studied and religious affiliation and the rest was available from Prolific (which is a platform used to administer the questionnaire as explained later):

- *Social class indicators.* We used three indicators of social class: Educational attainment and personal income as objective measures of social class and Subjective measure of one's rank in society which measures the *subjective* perception of one's rank in relation to others in society. For this, the MacArthur Scale of Subjective Socioeconomic Status [52] was deployed. Participants were asked to place themselves on a ladder with 10 rungs, which represent their position in society. For analysis purposes, the two topmost categories were merged due to very low numbers of participants locating themselves on rung 10, which indicated the highest possible social rank.
- *Subject of education.* We included education subject as a variable in our analysis to be able to examine the relationship between education and explanation preference from the angle of different fields of study rather than just the achieved level of education. This reflected the subject participants studied at their highest achieved level.
- *Gender.* In our study, gender was treated as a binary variable with levels female and male due to other categories (Genderqueer/Gender Non Conforming and those who preferred not to say) being too infrequent and potentially internally inconsistent to constitute its own category. Those who did not self-identify as female or male were excluded from statistical procedures testing gender, but their data where included in the rest of the tests. The terms *female* and *male* are used instead of *woman* and *man*, as those are the categories used by Prolific, the data collection platform used for this study.
- *Religious and Political beliefs.* Lastly, information about religious affiliation, political party affiliation and position on political spectrum are included. For context for those non-familiar with political parties in the UK, the relationship between parties and spectrum is represented in Figure 1 based on the data from our studv:

Fig. 1. Average position of UK political parties on a political spectrum derived from our data, plotted in Matplotlib [39].

4.3 Procedure

Participants for this study were recruited on Prolific (https://www.prolific.co) during May and June 2021 after running a pilot test with 38 participants.

We deployed a non-proportional quota sampling method [24] to ensure sufficient representation across the socio-cultural variables of interest, of gender, educational attainment, subjective perception of status, income, and political and religious beliefs.

		N	%			N	%
Gender	Female	299	65%	Religion	Non-religious	252	55%
	Male	154	33		Christian	163	35%
	Other/NA	7	2%		Other/NA	45	10%
Personal income	Less than £10,000	131	28%	Level	Secondary education (e.g. GED/GCSE)	68	15%
	£10,000 - £19,999	94	20%	of education	High school diploma/A-levels	87	19%
	£20,000 - £29,999	98	21%		Technical/community college	42	9%
	£30,000 - £39,999	38	8%		Undergraduate	175	38%
	Over £39,999	51	11%		Graduate	52	11%
	Other/NA	48	10%		Doctorate	28	6%
Subjective rank	1	30	7%		Other/NA	8	2%
in society	2	49	11%	Subject	Business & Administrative Studies	71	15%
	3	43	9%	of education	Creative Arts & Design	37	8%
	4	56	12%		Humanities, Languages and Edu	92	20%
	5	59	13%		Mathematical Sciences and Engineering	84	18%
	6	70	15%		Medicine and Life Sciences	51	11%
	7	69	15%		Psychology	26	6%
	8	53	12%		Social Sciences and Law	29	6%
	9+10	31	7%		Other/NA	70	15%
Political party	Conservative	73	16%	Political	Centre	154	33%
	Green Party	52	11%	spectrum	Left	201	44%
	Labour Party	156	34%		Right	55	12%
	Liberal Democrats	55	12%		Other/NA	50	11%
	SNP	28	6%				
	UKIP	30	7%				
	Other/NA	66	14%				

Table 2. Sample characteristics.

After providing informed consent, participants were presented with an information sheet that explains the decision-making process of BDI agents, referred to as 'robots' throughout the survey, in plain language, using a running example with a house cleaning robot. Then, they were shown the four scenarios in randomised order to avoid order effect. Each scenario begins with a short description of the agent, followed by instructions and questions for the *Preference as Explainee* part. Each participant was randomly assigned four out of eight questions per scenario. *Preference as Explainer* section followed in *shopping* and *pancake scenario*, comprising instructions and four questions per scenario in randomised order. Then, participants proceeded to the *cognitive style test* and *socio-cultural characteristics* section. The mean completion time was 17 minutes ($SD = 10\ min$) and participants were rewarded $2.16 for their contribution.

4.4 Participants

Originally, 503 participants were recruited, but only data from 460 participants made up the final data set ($N = 460$, 65% women and 1.5% Genderqueer/Gender Non Conforming; ($M_{age} = 38.2, SD_{age} = 13.5$). The remainder of 43 participants were excluded mostly for failing one or more attention checks (28 respondents) or as a result of not finishing the survey (15 respondents).

Due to the known effect of other factors on cognitive tendencies, such as language and nationality [4, 41], we restricted our sample to isolate the effect of our variables of interest, while trying to restrict the influence of the factors not considered in this study. Therefore, the inclusion criteria for the participants were (a) being nationals of the United Kingdom who (b) speak English as their first language and who (c) self-identify as monocultural. As shown in Table 2, this provided the diversity needed in terms of the socio-cultural factors of interest.

4.5 Data processing and analysis

This section describes the analytical tools used in the two parts of our study: Part 1 is concerned with the explanations participants prefer to receive as explainees and Part 2 studies which explanations participants prefer to give as explainers. These two parts operate with different dependent variables, described later in this section, but the independent variables are the same for both parts. The independent variables are gender, religion, personal income, subjective rank in society, political party affiliation, location on a political spectrum, level and subject of education and cognitive style. Cognitive style is represented by *Cognitive Style Index* (*CSI*, M=.50, SD=.37) which is a normalised additive index representing the proportion of analytic pairings chosen in the Triad Categorisation Task, where 0 means that the particular individual selected a holistic pairing in all test instances and 1 means that analytic pairings have always been selected.

Part 1 - Preference as Explainee. The dependent variable in Part 1 of the study is the *Explainee Preference Index* (*EPI*), capturing the preference for receiving either goal or belief explanation. The *EPI* (M=.51, SD=.21) is also a normalised additive index representing the proportion of belief explanations chosen out of all responses, which means that participants who preferred goal explanations for each instance of robot's action scored 0 and those always choosing belief explanations score 1 on the *EPI*.

In order to examine the relationship between the above-mentioned independent variables and *EPI*, (a) correlation analysis and (b) regression analysis were conducted. For the correlation analysis, a non-parametric Spearman's correlation was used because Shapiro-Wilks test indicated non-normal distribution of both CSI (W=.883, p<.001) and EPI (W=.983, p<.001), prohibiting the use of a parametric procedure. To correct for running multiple zero-order correlation tests between CSI, EPI and socio-cultural variables, Benjamini-Hochberg procedure with 10% false discovery rate was applied. 10% (or higher) false discovery rate is appropriate if missing potentially interesting findings is undesirable [55]. Because our study is the first of its kind, false negatives pose a greater threat than false positives, as potential false positives would be identified in future research. For the regression analysis, we adopted a fractional regression model, which is used to model the relationship between the independent variables and proportional discrete response variable. Due to all values of the outcome variable (*EPI*) falling within the unit interval $[0, 1]$, fractional regression is preferable to linear regression because linear regression can predict outside of the bounded interval and it is not able to account for non-linearity [29]. Another alternative to consider for modelling fractional outcome variables is beta regression, but beta regression is appropriate for variables within the open interval (0,1), excluding the endpoints [84]. Because our data contains both 0s and 1s for respondents with an absolute preference for goal explanations and belief explanations respectively, fractional regression is the appropriate choice. Fractional regression was introduced by Papke and Wooldridge [84] and it is a model estimated by Bernoulli quasi-maximum likelihood estimation (QMLE), typically with logit or probit response function and robust Huber-White sandwich error estimator. The estimates obtained by the Bernoulli QMLE are consistent and asymptotically normal irrespective of the distribution of the dependent variable and thus do not require any particular distribution [84].

These regression models serve the analytical purpose of identifying influential variables rather than predicting the outcome, which is neither desirable (see § 6.3) nor practical, as we only include specific socio-cultural factors and cognitive style without considering other potential predictors. As mentioned earlier, one of our objectives is to examine the relationship between social class and explanation preference. Because social class is a multidimensional concept, three indicators were used: educational attainment, income and a subjective measure of rank vis-a-vis society. These

three variables are intercorrelated, with the strength of correlation between medium and strong[3] [16], indicating that they partially capture the same concept, but they are also partially independent. Due to the intercorrelation, we only enter *Education level* into the fractional regression model, as it is the only social class indicator that is significantly associated with explanation preference according to the zero-order correlation (Table 3). To check for multicollinearity between the rest of the predictor variables, the variance inflation factor (VIF) was calculated. Considering all predictors $VIF_{mean} = 1.85$ and the highest detected VIF is for location on the political spectrum = Left ($VIF_{left} = 4.41$). Freund et al. [28] suggest that $VIF > 10$ indicates the presence of multicollinearity, others recommend caution for $VIF > 5$ [46]. Because all VIF values for our predictors are below five, we conclude that there is an absence of multicollinearity among our predictor variables.

To be able to investigate the mediating effect of cognitive style on the relationship between socio-cultural factors and explanation preference, two fractional regression models with a logit link were fitted: Model 1 comprising all socio-cultural factors and Model 2 which also contains *CSI*. The mediating effect of selected variables was also tested using the procedure proposed by Zhao, Lynch and Chen [110] using the Monte Carlo test [66] while including all of the other independent variables as covariates.

Part 2 - Preference as Explainer. Textual data for this part of the study were generated through the eight open-ended questions in which respondents were asked to explain actions performed by the BDI agents from the shopping and pancake scenario. In total, 3680 (460x8) free-text explanations were obtained and categorised according to the F.Ex Coding Scheme for People's Folk Explanations of Behavior[4]. The F.Ex scheme was developed by Malle in accordance with the folk-conceptual theory (§2.3) which is convenient for this study as it stems from the same intellectual tradition as the model of artificial BDI agents and, as such, includes goals and beliefs among other concepts, but also allows consideration of the *person-situation* distinction proposed by attribution theorists (§2.3). Even though the F.Ex coding offers more categories, the vast majority (98%) of our data were categorised as reasons, which makes sense, as we asked participants to explain the intentional behaviour of a robot. *Reasons* are mental states that the agent had when forming the intention to act. Reasons can be either goals[5], beliefs or valuings. From the data, 95% were classified as either belief or goal. Additionally, we also explored the content of beliefs people cited. That is, what the cited beliefs are about. 93% collected explanations were about either the agent itself or its situation. More context for the exploration of the belief content is available in § 5.2.2.

The explanations (N=3680) were coded by one author of this study and 10% of explanations (N=368) were coded by another researcher. The inter-rater agreement measured by Krippendorff's α is $\alpha = .80$ for explanation mode, $\alpha = .73$ for reason type, and $\alpha = .80$ for the content of belief explanations, which are considered to be acceptable values, suggesting good agreement [53]. Finally, two indices were calculated, which are the two dependent variables in Part 2 of this study. (a) *Explainer Preference Index* (*RPI*) is a normalised additive index, representing the fraction of belief (vs. goal) explanations made. Thus, people scoring 0 explained all robots' actions citing their goal and those scoring 1 always cited a belief when explaining the action of the robots. (b) *Belief Preference Index* (*BPI*) is also a normalised additive index calculated only from belief explanations and it represents the fraction of belief explanations with situation content (vs. agent content), where individuals always citing beliefs with agent content score 0 and those who made a belief

[3]Pairwise correlations coefficients: level of education and subjective rank in society (r=.34, p<.001), level of education and income (r=.40, p<.001), subjective rank in society and income (r=.47, p<.001)

[4]https://research.clps.brown.edu/SocCogSci/Coding/Fex%204.5.7%20(2014).pdf
https://research.clps.brown.edu/SocCogSci/Coding/fex.html

[5]F.Ex uses the term 'desires' rather than 'goals,' however, in this paper, we use 'goals' for consistency with Part 1 (§5.1)

explanation with a situation content in all instance score 1 on the *BPI*. Finally, a correlation analysis was performed using a non-parametric Spearman's correlation between the *RPI*, *BPI* and *EPI*, *CSI* and socio-cultural factors. To correct for running multiple zero-order correlation tests, the Benjamini-Hochberg procedure with a 10% false discovery rate was applied.

5 RESULTS

This section provides the results for the research questions set out in this study to help us describe the relationship between cognitive style, socio-cultural variables and explanation preference in BDI agents. Firstly, we present the results on explanation preference as explainee (recipient of explanation), collected using sets of closed questions in which respondents selected their preferred explanation per action of artificial BDI agent. Secondly, we report on the results regarding explanation preference from the perspective of the explainer (the producer of the explanation).

5.1 Part 1 - Preference as Explainee

5.1.1 Zero-order correlations. The first research question we aim to answer is concerned with the relationship between cognitive style and explanation preference.

Index→	Cognitive style CSI [Holistic→ Analytic] ρ (p)	Explanation preference EPI [Goal→ Belief] ρ (p)
Explanation preference index	.18 (<.001)	-
Gender = Female	...	-.15 (.001)
Gender = Male	...	.13 (.004)
Religion = Christian	...	-.15 (.002)
Religion = Non-religious	...	.15 (.002)
Income	.14 (.005)	...
Subjective status	.15 (.002)	...
Education level	.18 (<.001)	.15 (.001)
Subject: Business & Admin.	-.13 (.006)	-.12 (0.008)
- Creative Arts	...	...
- Humanities & Lang.	...	...
- Mathematical Sci	.13 (.007)	.12 (.010)
- Medicine & Life Sci	...	...
- Psychology	...	...
- Social Sci & Law	...	...
Party aff.: Conservative	...	...
- Green Party	.14 (.004)	.10 (.031)
- Labour Party	-.13 (.007)	-.11 (.021)
- Liberal Democrats	...	...
- SNP	...	...
- UKIP	...	...
Political aff.: Centre	...	...
- Left	...	...
- Right	...	...

Table 3. Statistically significant correlations (Spearman's ρ) between CSI and EPI with socio-cultural factors after correcting for multiple testing using Benjamini-Hochberg procedure with 10% false discovery rate. Ellipsis indicates non-significant correlations.

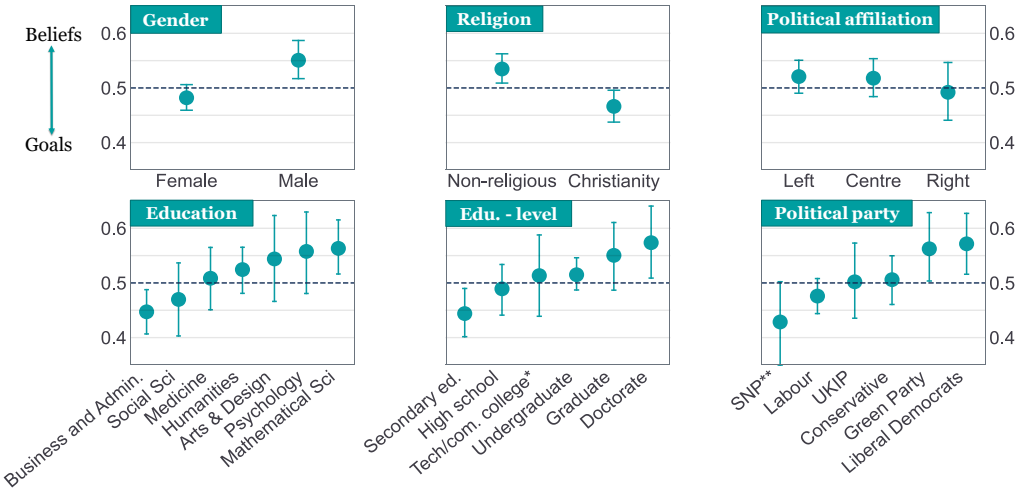

Fig. 2. Each graph represents the average Explanation Preference Index per category of socio-cultural factor. CI=95%.
*Technical or community college, **Scottish National Party.

Spearman's correlation indicates a positive association between EPI and CSI ($r(458) = .18$, $p<.001$). This means that *holistic* thinkers are associated with a preference for *goal* explanations and *analytic* thinkers are associated with *belief* explanations in BDI agent interaction (Table 3). The magnitude of the association is between small and medium [16]. Regarding the correlations between socio-cultural factors and explanation preference, being female ($r(458) = -.15$, $p=.001$), being Christian ($r(458) = -.15$, $p=.002$), having studied business and administration ($r(458) = -.12$, $p=.008$) and being a supporter of The Labour Party ($r(458) = -.11$, $p=.021$) is associated with a decrease in EPI, indicating a relatively higher preference for goal explanations. In contrast, being male ($r(458) = .13$, $p=.004$), being non religious ($r(458) = .15$, $p=.002$), achieving a higher level of education ($r(458) = .15$, $p=.001$), having studied mathematical sciences ($r(458) = .12$, $p=.010$) and being a supporter of The Green Party ($r(458) = .10$, $p=.031$) is associated with an increase in EPI, indicating a relatively higher preference for belief based explanations.

In terms of the association between cognitive style and socio-cultural factors, an important theoretical foundation of our study, we report that all three indicators of social class, income ($r(458) = .14$, $p=.005$), subjective perception of status ($r(458) = .15$, $p=.002$) and educational attainment ($r(458) = .18$, $p<.001$) are positively associated with CSI (Table 3), suggesting that individuals scoring higher on the social class scales are more analytic in thinking style while a lower score on the social class indicators is linked to holistic cognitive tendency, which corroborates the findings of [48, 51]. Our data also show the association between cognitive style and the subject of education, in the case of mathematical sciences ($r(458) = .13$, $p=.007$) and business and administration studies ($r(458) = -.13$, $p=.006$) which is consistent with [107] and some categories of political party affiliation. However, our results do not show a connection between gender, religion and cognitive style, contrary to the findings of [40, 80, 92], which could however be attributed to the idiosyncracies of the national contexts in which these studies have been conducted.

The results from the correlation analysis show us that there is a correlation between cognitive style and explanation preference, implying that *how one thinks* is to some degree related to whether

one prefers robots to explain themselves using goals or beliefs. However, from the correlation analysis, we can also observe that some socio-cultural categories are only correlated with either *Cognitive Style Index* or *Explainee Preference Index*, suggesting that for these categories, the differences in explanation preferences between socio-cultural factors are not mediated via cognitive style. However, in the case of educational attainment, some categories of education subject and political party affiliation, that are correlated with both the *CSI* and *EPI*, it is possible that the differences in explanation preference are mediated via cognitive style. The existence of the mediation effect of *CSI* is assessed in the multivariate regression analysis.

5.1.2 Multivariate regression. Next, we consider the socio-cultural factors and cognitive style together by entering them into a set of multivariate models in which gender, religiosity, education level, subject of education, political party affiliation, position on the political spectrum and cognitive style (excluded in Model 1) are the predictor variables and *EPI* is the outcome variable.

Table 4 summarises the average marginal effects (AME) of the two fractional regression models. AME indicates the average percentage point change in the outcome variable for a unit change in the predictor variable [77, 103]. Regression coefficients can be found in Supplementary Material 2.

In Table 4 we can observe that the results of Model 1, which does not include CSI, are fairly similar to the results of bivariate correlations between socio-cultural factors and explanation preference index. According to the fractional regression model, being a male (vs. female), having higher education attainment and having studied mathematical sciences (vs. business and administration)[6] are associated with an increase in EPI and thus show an increase in preference for belief explanations. Being a Christian (vs. non religious) and being a supporter of The Labour Party and The Scottish National Party (vs. Liberal Democrats) is associated with a decrease in EPI and as such indicates an increase in preference for goal explanation. Thus far, the results of the regression analysis corroborate our previous findings from the correlation analysis. There are, however, some differences, namely in subject of education and political party affiliation: according to the regression results, humanities and languages (vs. business and admin.) as a category of subject of education is associated with an increase in EPI, while being a supporter of The Green Party (vs. Liberal Democrats) is not associated with EPI at a level .05 of statistical significance. In addition to the investigation of main effects, we also tested all possible two-way interactions adding them to Model 2 one at a time. Most interaction terms did not have a statistically significant relationship with *EPI* ($\alpha = 0.05$). Only a very small subset of two-way interaction terms appeared statistically significant as detailed in Supplementary Material 3, but upon further investigation, the 2x2 cross-tabulation between the binary variables contained cells with low counts ($N_{cell} < 12$), suggesting that these results may be unreliable [9]. For this reason, further research with sufficient per-cell representation is required to investigate, whether any of these interaction terms are indeed significant.

Now, we turn to the comparison between Model 1 and Model 2 and thus examine the effect on the results of including CSI in the regression model. First, both categories of subject of education which were statistically significant (p<=.06) in Model 1 are no longer, which suggests that the variance explained by subject of education in Model 1 is explained by CSI in Model 2. The rest of the variables – gender, education attainment and religion and political affiliation — show very minimal differences when compared between Model 1 and 2, indicating independence of cognitive style index on those variables of socio-cultural background. In order to test whether *CSI* mediates the relation between being a graduate of *humanities & languages* or *mathematical sciences* and

[6]Note, the statistical significance of subject of education = Mathematical sciences is p=.06, which is beyond the traditionally accepted .05, but the authors deem it more important to report on potentially existing differences rather than honoring the arbitrary cut-off point.

Index→	**Explanation preference** Model 1 AME (SE)	**Explanation preference** Model 2 AME (SE)
CSI	X	0.06 (0.03)*
Gender: Male	0.05 (0.022)*	0.05 (0.02)*
- *DK/NA (N=7)*	...	...
Religion: Christian	-0.04 (0.021)*	-0.05 (0.02)*
- DK/NA	...	...
Education level	0.02 (0.006)**	0.02 (0.01)*
Subject: Creative Arts	...	...
- Humanities & Lang.	0.06 (0.029)*	...
- Mathematical Sc's	*0.06 (0.03)*°	...
- Medicine and Life Sc's	...	...
- Psychology	...	...
- Social Sc's and Law	...	...
- Other	...	...
Pol. party: Conservative	...	...
- Green Party	...	...
- Labour Party	-0.07 (0.032)*	-0.07 (0.03)*
- SNP	-0.11 (0.045)**	-0.12 (0.05)**
- UKIP	...	...
- DK/NA	...	...
Pol. spectrum: Centre	...	...
- Left	...	...
- DN/NA	...	...
Pseudo R^2	0.017	0.019

Table 4. Summary of multivariate logistic regression models, showing average marginal effects (AME) for variables statistically significant at level 0.05. Marginal effects indicate the percentage point change in the outcome variable for a unit change in the predictor variable. Baseline values are gender[female], religion[non-religious], subject[Business and administrative studies], political party affiliation[Liberal Democrats], political affiliation[right]. Ellipsis indicates non-significant result, *X* indicates that variable is not included in the model. °p=.06, *p<.05, **p<.01.

EPI, we used the procedure proposed by Zhao, Lynch and Chen [110] using the Monte Carlo test [66] while including all of the other independent variables as covariates. For both of the tested independent variables, *humanities & languages* and *mathematical sciences*, the Monte Carlo tests were not significant ($p_{humanities} = 0.92$, $p_{math} = 0.10$), which indicates that cognitive style does not mediate the relationship between being graduate of *humanities & languages* or *mathematical sciences* and *EPI*. Since the rest of the predictors perform the same in both models, we conclude that cognitive style does not mediate the relationship between any socio-cultural factor included in our study and *EPI*.

These findings (Model 2) corroborate those from the correlation analysis regarding cognitive style, represented by the *CSI*, suggesting that *how* one thinks is associated with goal or belief explanation preference. However, the effect of cognitive style is independent of the effect of gender, religiosity, education level, and political affiliation, which means that there is another mechanism through which these socio-cultural factors relate to explanation preference than cognitive style. These findings thus prompt further research to understand the mechanism via which these socio-cultural categories are associated with differences in explanation preferences in BDI agents.

	Explainer Preference Index (*RPI*) [Goal→ Belief] *ρ (p)*
EPI [Goal→ Belief]	.39 (<.001)
CSI [Holistic→ Analytic]	.18 (<.001)
Gender: Female	-.13 (.004)
- Male	.11 (.015)
Religion: Christianity	-.13 (.006)
- Non-religious	.11 (.017)
Educational attainment	.14 (.002)
Subjective status	...
Income	...
Subject: Business & admin.	...
- Arts & design	...
- Humanities & lang.	...
- Mathematical sci.	...
- Medicine & life sci.	...
- Psychology	...
- Social sci.	...

Table 5. Zero-order correlations (Spearman's ρ) between *Explainer Preference Index* and *Explainee Preference Index*, *Cognitive Style Index* and socio-cultural factors. Political party affiliation and political affiliation are omitted due to non-significant results for any category. Statistically significant correlations after correcting for multiple testing using Benjamini-Hochberg procedure with 10% false discovery rate are shown. Ellipsis indicates non-significant correlations.

5.2 Part 2 - Preference as Explainer

So far, we have examined the preference participants expressed from the position of an explainee. Here we report on participants' preference as an explainer – that is, how they explain the behaviour of artificial BDI agents. The aim is to observe the association between cognitive style, socio-cultural background and explanation type made by the participants rather than letting them choose from a restricted set of explanations. This section first describes the findings regarding the split between goals and beliefs and an exploration into the content of the beliefs.

5.2.1 Goals vs Beliefs. Here we inspect the tendency to give either goal or belief explanations (*RPI*) and its relation to cognitive style, socio-cultural factors, but also with *EPI*— the index capturing preference as explainee — in order to test if people who prefer to receive a certain type of explanation are also more likely to offer an explanation of the same kind. Below are examples of explanations coded as beliefs and goals:

Question: Why did you (the pancake-making robot) add sugar to the bowl?

- *Goal:* "I wanted all the ingredients in the bowl so that I can prepare the batter." (Ex. 901)
- *Belief:* "Because the recipe stated sugar was needed and the sugar was not in the bowl." (Ex. 821)

Question: Why did you (the shopping robot) decide to shop online?

- *Goal:* "[I] wanted to shop quickly without leaving the house." (Ex. 2504)
- *Belief:* "Because the physical shop was closed at that time." (Ex. 2673)

In terms of the relationship between explanation preference as explainee (*EPI*) and explainer (*RPI*), our data suggest a positive association according to Spearman's correlation (r(458)=.39,

p<.001) meaning that those who prefer to receive more goal explanations are more likely to explain behaviour by goal explanations and vice versa. *Cognitive Style Index* and *RPI* are also positively associated (r(458)=.18, p<.001), suggesting that the more holistic one's cognitive style, the more of goal explanations they make and with increase in analytic cognitive style, belief explanations are more frequent. Our data also shows an association between *RPI* and gender, religion affiliation, and educational attainment. Being a male (r(458)=.11, p=.015), non-religious (r(458)=.11, p=.017) and scoring higher on the educational attainment measure (r(458)=.14, p=.002) is associated with higher *RPI*, hence a stronger preference for explaining behaviour of BDI agents by citing beliefs. Being a female (r(458)=-.13, p=.004) and being Christian (r(458)=-.13, p=.006) is associated with lower *RPI*, suggesting that females and Christians have a higher tendency to mention goals as explanations. No association was detected between *RPI* and political affiliation. Table 5 summarises correlation coefficients between *RPI* and *EPI*, *CSI*, and socio-cultural factors for those correlations that are statistically significant after correcting for multiple testing using the Benjamini-Hochberg procedure with 10% false discovery rate [5, 100].

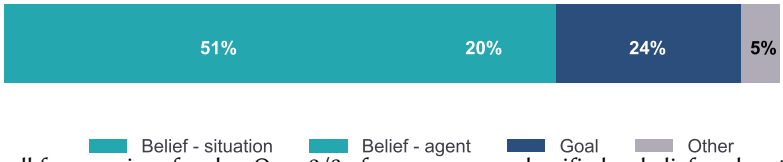

Fig. 3. Overall frequencies of codes. Over 2/3 of answers were classified as belief explanation with belief - situation being the most common explanation: more than half of all answers were coded as belief - situation explanation.

5.2.2 Belief content (agent vs. situation). Although there is a significant correlation between *RPI*, *CSI* and socio-cultural factors, we observed a key difference with respect to preferences as explainees, which is that there are more belief than goal explanations generated by open-ended questions. As we can see in Figure 3, there were many more belief explanations than goal explanations. This contrasts with preference as explainee, where there was a roughly balanced split between preferences for goal and belief explanations (Mean *EPI* = .51). The dominant preference of explainers for belief explanations leads us to explore if there is any systematic variation between the kinds of beliefs people cite in relation to cognitive style and socio-cultural factors of interest to this study. Driven by the theoretical underpinning of this study, attribution theory in particular, and aided by the F.Ex coding scheme, we further categorised beliefs according to their *content*; that is *what* is believed in order to test the person-situation distinction. Beliefs could be held about the situation in which the agent finds itself (situation content) or about the agent itself (agent content). As shown in Figure 3, the majority (67%) of belief explanations were classified as having situation content, 26% of belief explanations were about the agent and 6% of belief explanations were categorised as other.

Below are examples of belief explanations with either agent or situation content:

Question: Why did you (the pancake-making robot) throw away the first pancake?

- *Agent content:* "I burnt it". (Ex. 993)
- *Situation content:* "The pan was not hot enough yet and so the first pancake cooked oddly". (Ex. 979)

Question: Why did you (the shopping robot) take a bus?

- *Agent content:* "I do not drive". (Ex. 3632)
- *Situation content:* "Because the shop was too far to walk to, especially with heavy bags on the way home". (Ex. 3678)

	Belief Preference Index (*BPI*) [Agent→ Situation] *ρ (p)*
EPI [Goal→ Belief]	.31 (<.001)
CSI [Holistic→ Analytic]	.15 (.001)
Gender: Female	-.16 (<.001)
- Male	.17 (<.001)
Religion: Christianity	...
- Non-religious	...
Educational attainment	.17 (<.001)
Subjective status	...
Income	...
Subject: Business & admin.	-.10 (.025)
- Arts & design	...
- Humanities & lang.	...
- Mathematical sci.	...
- Medicine & life sci.	...
- Psychology	...
- Social sci.	...

Table 6. Zero-order correlations (Spearman's ρ) between *Belief Preference Index* and *Explainee Preference Index*, *Cognitive Style Index* and socio-cultural factors. Political party affiliation and political affiliation are omitted due to non-significant results for any category. Statistically significant correlations after correcting for multiple testing using Benjamini-Hochberg procedure with a 10% false discovery rate are shown. Ellipsis indicates non-significant correlations.

We examined the relationship between *BPI*, *EPI*, *CSI*, and socio-cultural factors through a set of Spearman's correlations. We found a positive correlation between *BPI* and *EPI* (r(458)=.31, p<.001), suggesting an association between preference for goal explanation and agent content in belief explanations and an association between belief explanations and situation content. Cognitive style is also found to be related to the content of the belief explanation. *CSI* and *BPI* are positively correlated (r(458)=.15, p<.001) which means that holistic thinkers are more likely to refer to agents, and analytic thinkers tend to cite situations when making belief explanations. From the socio-cultural factors, *BPI* is associated with gender, educational attainment and some categories of a subject of education. Indicated by a positive correlation, being a male (r(458)=.17, p<.001) and having higher educational attainment (r(458)=.17, p<.001) is associated with a stronger preference for situation content of beliefs. Being a female (r(458)=-.16, p<.001) and having studied business and administration studies (r(458)=-.10, p=.025) is associated with a decrease in *BPI*, hence suggesting an increase in the tendency to make belief explanations with agent content. Religious and political affiliation is not associated with a preference for belief explanation content as suggested by our data. Table 6 summarises statistically significant correlations after correcting for multiple testing using Benjamini-Hochberg procedure with a 10% false discovery rate.

In terms of the goal vs. belief distinction, the results show us that as well as explainee's preference, explainer's preference is associated with cognitive style and also that the stronger people prefer a particular explanation as explainees, the stronger they tend to prefer the same explanation as explainers. This becomes apparent also in the correspondence of socio-cultural categories significantly associated with *RPI* – gender, religiosity and educational attainment – as those are also categories exhibiting differences in *EPI*. Interestingly, except for religiosity, these variables are also associated with differences in belief content for explainees, indicating that those categories

			Preference as Explainee (Part 1) Explanation	**Preference as Explainer** (Part 2) Explanation	Belief Content
	Preference as Explainee	Goals	–	Goals	Agents
		Beliefs	–	Beliefs	Situation
H1	Cognitive Style	Holistic	Goals	Goals	Agents
		Analytic	Beliefs	Beliefs	Situation
H2.1	Education Level	Low	Goals	Goals	Agents
		High	Beliefs	Beliefs	Situation
H2.2	Subject of Education	Business and admin.*	Goals	…	Agents
		Mathematical Sci.*	Beliefs	…	…
H2.3	Gender	Female	Goals	Goals	Agents
		Male	Beliefs	Beliefs	Situation
H2.4	Religion	Christian	Goals	Goals	…
		Non-religious	Beliefs	Beliefs	…
H2.5	Pol. spectrum	…	…	…	…
H2.6	Political party	Labour	Goals	…	…
		Green Party*	Beliefs	…	…

Table 7. Summary of results from the set of bivariate correlations. The column Preference as Explainee summarises findings from § 5.1, column Preference as Explainer - Explanation summarises findings from § 5.2.1 and column Preferences as Explainer - Belief Content summarises findings from § 5.2.2. * marks categories that are not significant in the fractional regression model (Model 2). Cognitive style was not found to mediate the relationship between any independent variable and explanation preference and it is not represented in the table for the sake of clarity.

associated with differences in goal vs. belief preferences are also associated with differences in belief content.

6 DISCUSSION

6.1 Summary of findings

To offer an integrated perspective of our findings, Table 7 shows a summary of the bivariate correlations from both parts of the study. Part 1 details explainee explanation preference (§ 5.1) and Part 2 presents explainer explanation preference (§ 5.2). With regards to our research questions:

RQ1: What is the relationship between cognitive style and preference for explanation type in BDI agents? Our results suggest that cognitive style and preference for explanation preference in BDI agents are significantly associated. In particular, holistic thinkers have a higher preference for goal explanation as both explainees and explainers and, in the case of belief explanation, they have a higher preference for agent content. Analytic thinkers have an increased preference for belief explanations, as both explainees and explainers and for belief explanations with situation content in particular.

RQ2: What is the relationship between socio-cultural factors and explanation type in BDI agents? Is it mediated via cognitive style? Despite some minor differences, our results suggest that gender, religion, educational attainment, the subject of one's education and political party affiliation all affect what reasons people prefer in explainations of intentional behaviour, in our case the actions of artificial BDI agents. Our data suggest that females, Christians and people with lower educational attainment are more likely to prefer goal explanations as both explainers and explainees. In contrast, male, non-religious people and those with higher education level tend to prefer belief

explanation. Some categories of education subject and political party affiliation are also connected with explanation preference, though only as an explainee. Graduates of business and administration studies and supporters of The Labour Party prefer to receive goal explanations, while graduates of mathematical sciences and supporters of The Green Party prefer belief explanations. The effect of cognitive style is mostly independent of the effect of the socio-cultural variables.

6.2 Structure of the mind matters

Our results show that cognitive style and socio-cultural factors are indeed associated with the user's explanation preferences in BDI agents, which would, in principle, be consistent with the canonical literature on cross-cultural differences in cognitive style and causal attribution [49, 73, 76]. However, the orientation of this association is the *opposite* from what that literature suggests. According to the existing literature, holistic thinkers tend to make external attribution and analytic thinkers are more likely to make internal attribution. In contrast, in our data and particularly in belief explanations, holistic thinkers tended to cite beliefs *about the agent*, thus making an internal attribution, while analytic thinkers more often provided belief *about the situation* and thus making an external attribution.

Let us instead view the results with folk-conceptual lenses. As far as we know, there is currently no available literature on cross-cultural differences in the use of beliefs and goals, but there is evidence for the actor-observer asymmetry in the use of reasons, which is caused by differences in access to information and motivation of the explainer. The difference between explanations of actors and observers is based on access to information, and motivation to attribute rationality to the agent, and is linked to the different properties of goals and beliefs [61]. This might offer a useful starting point for discussing our results by considering how these factors could relate to the different groups of people delimited by socio-cultural characteristics, instilling the disposition to prefer a particular explanation type [61].

The actor-observer asymmetry in the use of beliefs or goals as explanations lies in the tendency to provide a belief reason for own behaviour and a goal explanation for the behaviour of another person [61]. According to Malle, *goals* (desires in their work) are easier to infer from behaviour because they are often informed by context and cultural scripts and they can be often revealed by behaviour. *Goals* also represent a wanting or lacking [59], which is more generic [61] and thus cognitively cheaper. *Beliefs* in comparison are the result of a more complex cognitive process and stem from the interaction between an agent's knowledge, outcome prediction, contextual information and their causal relationship. Beliefs are more idiosyncratic in nature and free of the societal constraints to which goals are subjected, and so are harder to infer, especially if the observer does not share context with the actor [59]. This is supported by the empirical evidence of the actor-observer asymmetry: people are more likely to cite beliefs when explaining their behaviour (actors), as they have access to idiosyncratic information resulting from more complex consideration, but people (observers) are more likely to provide goals to explain the behaviour of others. This might be because they do not have access to the idiosyncratic and hard-to-infer beliefs of the actor, or because they lack the motivation to portray the actor as a rational, thinking agent rather than merely a wanting actor [61]. These findings are not directly relevant to our results, as we operate only with observers, but the underlying mechanisms generating the differences between actors and observers based on information access and motivation to portray the agent as rational might inspire some hypotheses for the existence of differences between socio-cultural categories, stemming from different cultural, attitudinal and cognitive characteristic. Thus while the differences in using the attribution-theoretical person-situation distinction do not seem to provide a possible interpretation of the results from our study, we propose three possible explanations derived from folk-conceptual theory. These will have to be scrutinised in further research:

1) Social scripts. From the work of Malle et al. [59, 61], we learn that culture puts bounds on goals and that they are embedded in cultural schemata and scripts, while beliefs are more specific to individuals. This interpretation suggests that our initial proposal regarding the link between cognitive styles and explanation type based on the notion that beliefs are external and preferred by holistic thinkers, while goals are internal and preferred by analytic thinkers is false. Recall that cognitive style is believed by some [102] to be predicated upon the social orientation of culture, hence whether one has independent (individualist cultures) or interdependent (collectivist cultures) self-construal. This interpretation then suggests an alternative hypothesis: the reason holistic thinkers prefer goal explanations is because of their embeddedness in the culturally shared set of practices, meanings, scripts and constraints. Analytic thinkers, who are typical for their independent self-construal and background in individualist cultures, favour belief explanation, which provides a specific, individualist account stemming from the actor's knowledge, intuitions, outcome predictions and judgements [59].

2) Motivation to grant rationality to robots. Although there is evidence that people tend to treat AI agents as entities endowed with human-like qualities, such as mental states, and personalities [20, 98] and so it seems reasonable to assume that people would enact their explanatory tendencies in explaining the behaviour of AI agents similarly as they would reason about the behaviour of humans, there is also evidence of some distinctions. De Graaf and Malle [21] show that even when people use the same concepts to describe robots' as well as humans' behaviour, there are some differences, for example in the proportion between the use of goals and beliefs, which suggests the existence of unique factors that come into play in explanation of robots' behaviour. Here, we propose that one of these factors could be the humans' attitude toward robots influencing their willingness to grant rationality to the robot.

Remember that the decision about whether to explain behaviour with a goal or a belief hinges on two factors: the availability of the information and the *motivation to portray the actor as rational* [61]. Hence, the results might suggest that some subpopulations are more inclined to perceive AI agents as rational entities rather than primitive machines merely pursuing their goals. Attitudes toward service robots vary according to gender and occupation, where women and people working in non-social careers (which might be related to the subject of education) had more negative attitudes towards service robots [87, 91]. This might hinder their willingness to attribute rationality to the agents. There is mixed evidence on the influence of education on attitude towards robots [87, 91].

De Graaf and Malle [21], however, offer another related perspective on the perception of rationality in robots. In their study, they found that while people offer more goal explanations for the actions of fellow humans, they offer a similar number of goal and belief explanations for robots, thus citing relatively more belief explanations for robots. Their interpretation of this finding is that people are more comfortable seeing robots are rational entities with beliefs rather than affective entities with desires. This is because knowing only the goal opens a space in which the goal can be understood both as a result of rational deliberation or arational desire [7], while beliefs may be more closely connected to rationality. In the context of our study, it could thus be hypothesised that the differences observed in our study are due to the different perceptions of rationality and the affectivity of robots between the subpopulations in our study.

3) Willingness to exert cognitive effort. Another hypothesis might be related to the need for cognition (NfC), which is an individual's inclination to engage in thinking [10]. The results might suggest that those subpopulations who tend to prefer beliefs do so because they seek the information, which is less readily available and requires more cognitive effort, while those with a low need for cognition are satisfied with goal explanations, which are easier to infer and thus require less

[7] Arational desires are not a result of conscious deliberation, such as an arational desire for food [90].

cognitive effort. Previous research does not support this hypothesis unequivocally: even though there is evidence of the differences in NfC between professions, which could be a proxy for our subject of education, education and cognitive style, gender differences have not been found [10, 93] hence this hypothesis is relatively weak.

To summarise, folk-conceptual theory provides us with information regarding the properties of goals and beliefs (their level of difficulty to infer from action alone, cultural and societal constraints) [59] and the factors influencing the tendency of people to use these reasons as explanations (availability of information and motivation to portray actors as rational), that are manifested in the actor-observer asymmetry [61]. Equipped with this information, we formulated three hypothetical explanations of our empirical results that will have to be tested in future research.

6.3 Recommendations to the design community

We advise against considering our findings as justification for the use of the socio-cultural characteristics of the user to offer personalised explanations because using group-level data as a basis for making predictions about the preferences of individuals might result in harmful, discriminatory treatment of those whose personal preferences differ from the majority of people belonging in the same socio-cultural categories [27, 101]. The use of socio-cultural factors in selecting an explanation could also cause privacy issues [95] and, if insufficiently calibrated, could lead to AI systems harmfully discriminating and amplifying stereotypes, more so in that it would be of no practical importance to our study, as all of our predictors are relatively weak, even though they show clear associations. Furthermore, the categorisation of users into categories based on socio-cultural factors in itself can be problematic in any context if the representation of the categories is (a) incomplete, such as the binary representation of gender, or (b) immutable, because users' belonging to certain categories, for example, political or religious affiliation or educational attainment, can change in the course of one's life. Additionally, (c) disregarding intersectionality of the user's identity can result in mistreatment when the interactions of different facets of the user's identity are not recognised. This tension between the provision of generic explanations and thus ignoring the group-level evidence on different preferences and the provision of 'personalised' explanations based on inferring individual-level preferences from group-level data opens up an interesting debate about whether equality or equity should be striven for. However, such a discussion is out of the scope of this paper.

As a compromise between personalisation of explanation and providing one kind of explanation to all users, *our recommendation to the UX design community is to co-create explanations with users, that contain both belief and goal* to cater to users with a preference for either. This recommendation is in agreement with the recommendation of Harbers et al. [34]. However because there is evidence of a preference for simple explanations over more complex ones [56, 85], we offer our recommendations as tentative until further research provides us with a better understanding of the mechanism behind the observed differences, which might enable us to offer less problematic personalised explanations. This leads to our second recommendation, which is aimed at the research community. *Further research shall be conducted to understand the underlying mechanism that drives the observed differences along several socio-cultural dimensions.* This is because membership in a socio-cultural category is not a reason for a different preference in itself, but it can be a proxy, however a noisy one, for some underlying mechanism driving these differences. If we manage to identify what drives these differences, then we can aim to classify individual users in relation to the underlying mechanism and provide them with a personalised explanation which would enhance understanding while avoiding the presentation of redundant information. On a more abstract level, our final recommendation to the research and design community is the following: *in light of our results, we emphasise the importance of testing and validating all systems (not only BDI-based) with the*

intended users to ensure it is their preferences for which the system is optimised. The importance of conducting user studies with the real users [19] and accounting for the diversity among them [81] is well documented in the literature, and by identifying different preferences based on cognitive styles and socio-cultural factors, our study offers further perspectives to consider for that diversity, particularly in the case of BDI agents.

In summary, we believe that our findings are important in three ways. First, we corroborate the suggestions of Harbers [34] in constructing BDI agent explanation. Second, we set a stepping stone into the exploration of the relationship between culture and explanation preferences in AI and finally, our findings emphasise the importance of well-targeted user studies.

6.4 Limitations and future research

We acknowledge some limitations of this study stemming from its scope and methodology and we also propose some directions for future research. In order to isolate the effect of the socio-cultural factors selected in this study (gender, social class, subject of education and religious and political affiliation), the pool of participants has been restricted in terms of nationality and language, which are also factors known to influence cognitive processes [4, 41]. The findings should therefore not be generalised beyond this context. Future research is suggested to include those additional factors and/or to be conducted in a different linguistic and national context. Follow-up qualitative studies could also help contextualise and better understand the differences observed in this present study, especially if conducted in more realistic settings, as using proxy tasks instead of observing realistic human-AI interactions is known to impact the results [8]. This study includes four different scenarios balancing the ones familiar and unfamiliar to the user across different topics in order to cover a range of situations as done in previous work [34]. Despite this effort, another possible limitation, which is left unexplored in this study, is the impact of the level of expertise and the specific scenarios on the results. In particular, the level of expertise in a given domain has been found to be an important user factor influencing human-AI interactions and explanation needs. Domain expertise is an important factor influencing trust in XAI system when encountering errors [78] and even in the context of goal and belief explanations in BDI agents, it has been suggested that the user's domain expertise might be a distinctive factor in explanation preference [34]. Apart from different explanation preferences and attitudes towards the system, domain expertise, or the lack thereof, can be also problematic if the user simply does not understand the explanation due to the lack of domain-specific knowledge. In this case, a promising solution was proposed by He et al. [36] to use an analogy-based explanation, which expands concept-level explanation with an analogy to help users understand the explanation. An interesting avenue for future research would be to investigate the usefulness of analogy-based extensions of belief and desire explanations in unknown domains.

Another exciting and important line of research would be to investigate more human factors in terms of belief or goal explanation preference. In this paper, we investigated selected socio-cultural factors and cognitive styles in relation to goal or belief explanation, but other user characteristics are relevant to some aspects of human-AI interactions, such as attitude towards risk, computer self-efficacy, motivations, information processing style, and learning style [2], cognitive biases [37], reading proficiency, conscientiousness and need for cognition [17].

At the beginning of this paper, we acknowledged that explanations are both cognitive and social processes [69] and focussed on the cognitive aspect of explanations. However, it is crucial to understand also the social element, that is how best to communicate explanations and what are the implications of using different modalities of explanations [89].

Finally, another limitation is that by choosing BDI agents, we limited ourselves to studying the difference between goal and belief explanations. However, there could be more pronounced differences along different dimensions beyond the belief-goal-based distinction.

7 CONCLUSION

This paper contributes to the scholarship on individual and cultural differences in human-agent interaction by identifying factors associated with explanation preference, such as cognitive style and some socio-cultural factors, and it provides recommendations to AI researchers and designers based on the findings. In particular, this paper studied the relationship between socio-cultural background, cognitive style and preference for belief or goal explanation in artificial BDI agents, where participants were both on the receiving and on the giving end of the explanation for the artificial agent's action. The main findings of this study are that both cognitive style and some socio-cultural factors are associated with a preference for a goal or belief explanations, but the effect of cognitive style and socio-cultural factors is independent. Our data suggest that analytic thinkers prefer belief explanations and people with holistic cognitive style prefer goal explanations. Socio-cultural variables that affect explanation preference are gender, religiosity, educational attainment, some fields of education, and political party affiliation. Position on the political spectrum has not been found to affect explanation preference.

ACKNOWLEDGMENTS

We would like to thank anonymous reviewers, Alfie Abdul Rahman, Xiao Zhan and Richard Willis for their comments on the previous versions of this manuscript, Sara Tandon for her feedback on the questionnaire used in this study and Elfia Bezou-Vrakatseli for coding a portion of the qualitative data. This work was supported by UK Research and Innovation [grant number EP/S023356/1], in the UKRI Centre for Doctoral Training in Safe and Trusted Artificial Intelligence (www.safeandtrustedai.org). This research was partially funded by the INCIBE's strategic SPRINT (Seguridad y Privacidad en Sistemas con Inteligencia Artificial) C063/23 project with funds from the EU-NextGenerationEU through the Spanish government's Plan de Recuperación, Transformación y Resiliencia.

REFERENCES

[1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y Lim, and Mohan Kankanhalli. 2018. Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. Association for Computing Machinery, New York, NY, USA, 1–18.

[2] Andrew Anderson, Tianyi Li, Mihaela Vorvoreanu, and Margaret Burnett. 2021. Diverse Humans and Human-AI Interaction: What Cognitive Style Disaggregation Reveals. 99, 99 (2021). arXiv:2108.00588 http://arxiv.org/abs/2108.00588

[3] Andrea Bender and Sieghard Beller. 2013. Cognition is … fundamentally cultural. *Behavioral Sciences* 3, 1 (2013), 42–54. https://doi.org/10.3390/bs3010042

[4] Andrea Bender, Sieghard Beller, and Douglas L. Medin. 2017. Causal Cognition and Culture. In *Oxford Handbook of Causal Reasoning*, Michael R. Waldmann (Ed.). Oxford University Press, New York, USA, 717–38. https://doi.org/10.1093/oxfordhb/9780199399550.013.34

[5] Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)* 57, 1 (1995), 289–300.

[6] Michael Bratman. 1987. *Intention, plans, and practical reason.* Harvard University Press, Cambridge, Mass.

[7] Jill Brown, Colin A. McDonald, and Fabiola Roman. 2014. "The dog and the carrot are both useful to me": Functional, self-referent categorization in rural contexts of scarcity in the Dominican Republic. *International Perspectives in Psychology: Research, Practice, Consultation* 3, 2 (2014), 63–75. https://doi.org/10.1037/ipp0000014

[8] Zana Buçinca, Phoebe Lin, Krzysztof Z. Gajos, and Elena L. Glassman. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. *International Conference on Intelligent User Interfaces, Proceedings IUI* (2020), 454–464. https://doi.org/10.1145/3377325.3377498 arXiv:2001.08298

[9] Katherine S. Button, John P.A. Ioannidis, Claire Mokrysz, Brian A. Nosek, Jonathan Flint, Emma S.J. Robinson, and Marcus R. Munafò. 2013. Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience* 14, 5 (2013), 365–376. https://doi.org/10.1038/nrn3475

[10] John T. Cacioppo and Richard E. Petty. 1982. The need for cognition. *Journal of Personality and Social Psychology* 42 (1982), 116–131.

[11] Yuqin Chen, Yuanjun Dai, and Yufen Liu. 2021. Design & Implementation of Airport Runway Robot Based on Artificial Intelligence. In *2021 IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, Vol. 5. IEEE, 2636–2640. https://doi.org/10.1109/IAEAC50856.2021.9390864

[12] Incheol Choi, Minkyung Koo, and Jong An Choi. 2007. Individual differences in analytic versus holistic thinking. *Personality and Social Psychology Bulletin* 33, 5 (2007), 691–705. https://doi.org/10.1177/0146167206298568

[13] Incheol Choi, Richard E. Nisbett, and Ara Norenzayan. 1999. Causal Attribution Across Cultures: Variation and Universality. *Psychological Bulletin* 125, 1 (1999), 47–63. https://doi.org/10.1037/0033-2909.125.1.47

[14] Jaewon Choi, Hong Joo Lee, Farhana Sajjad, and Habin Lee. 2014. The influence of national culture on the attitude towards mobile recommender systems. *Technological Forecasting and Social Change* 86 (2014), 65–79. https://doi.org/10.1016/j.techfore.2013.08.012

[15] Nazli Cila. 2022. Designing Human-Agent Collaborations: Commitment, Responsiveness, and Support. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 420, 18 pages. https://doi.org/10.1145/3491102.3517500

[16] J. Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, Hillsdale, NJ, USA.

[17] Cristina Conati, Oswald Barral, Vanessa Putnam, and Lea Rieger. 2021. Toward personalized XAI: A case study in intelligent tutoring systems. *Artificial Intelligence* 298 (2021), 103503. https://doi.org/10.1016/j.artint.2021.103503

[18] Enrico Costanza, Joel E. Fischer, James A. Colley, Tom Rodden, Sarvapali D. Ramchurn, and Nicholas R. Jennings. 2014. Doing the Laundry with Agents: A Field Trial of a Future Smart Energy System in the Home. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 813–822. https://doi.org/10.1145/2556288.2557167

[19] Catherine Courage and Kathy Baxter. 2005. *Understanding your users: A practical guide to user requirements methods, tools, and techniques*. Gulf Professional Publishing.

[20] Maartje M.A. De Graaf and Bertram F. Malle. 2017. How people explain action (and autonomous intelligent systems should too). *AAAI Fall Symposium - Technical Report* FS-17-01 - FS-17-05 (2017), 19–26.

[21] Maartje M.A. De Graaf and Bertram F. Malle. 2019. People's Explanations of Robot Behavior Subtly Reveal Mental State Inferences. *ACM/IEEE International Conference on Human-Robot Interaction* 2019-March (2019), 239–248. https://doi.org/10.1109/HRI.2019.8673308

[22] Upol Ehsan, Samir Passi, Qingzi Vera Liao, Larry Chan, I-Hsiang Lee, Michael J. Muller, and Mark O. Riedl. 2021. The Who in Explainable AI: How AI Background Shapes Perceptions of AI Explanations. *ArXiv* abs/2107.13509 (2021), 47 pages.

[23] Connor Esterwood, Kyle Essenmacher, Han Yang, Fanpan Zeng, and Lionel Peter Robert. 2021. A Meta-Analysis of Human Personality and Robot Acceptance in Human-Robot Interaction. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 711, 18 pages. https://doi.org/10.1145/3411764.3445542

[24] Ilker Etikan and Kabiru Bala. 2017. Sampling and Sampling Methods. *Biometrics & Biostatistics International Journal* 5, 6 (2017), 215–217. https://doi.org/10.15406/bbij.2017.05.00149

[25] Anthony Faiola and Karl F. MacDorman. 2008. The influence of holistic and analytic cognitive styles on online information design: Toward a communication theory of cultural cognitive design. *Information Communication and Society* 11, 3 (2008), 348–374. https://doi.org/10.1080/13691180802025418

[26] Juliana J Ferreira and Mateus S Monteiro. 2020. What Are People Doing About XAI User Experience? A Survey on AI Explainability Research and Practice. In *Design, User Experience, and Usability. Design for Contemporary Interactive Environments*, Aaron Marcus and Elizabeth Rosenzweig (Eds.). Springer International Publishing, Cham, 56–73.

[27] Xavier Ferrer, Tom van Nuenen, Jose Such, Mark Cote, and Natalia Criado. 2021. Bias and Discrimination in AI: a cross-disciplinary perspective. *IEEE Technology and Society* 20, 2 (2021), 72–80.

[28] Rudolf J. (Rudolf Jakob) Freund, William J. Wilson, and Ping. Sa. 2006. *Regression analysis : statistical modeling of a response variable*. (2nd ed. / rudolf j. freund, william j. wilson, ping sa. ed.). Elsevier Academic Press, Burlington, MA.

[29] Susanna Gallani, Jeffrey M Wooldridge, and Ranjani Krishnan. 2015. Applications of Fractional Response Model to the Study of Bounded Dependent Variables in Accounting Research.

[30] Sorin Grigorescu, Bogdan Trasnea, Tiberiu Cocias, and Gigel Macesanu. 2020. A survey of deep learning techniques for autonomous driving. *Journal of Field Robotics* 37, 3 (2020), 362–386.

[31] Igor Grossmann and Michael E.W. Varnum. 2011. Social class, culture, and cognition. *Social Psychological and Personality Science* 2, 1 (2011), 81–89. https://doi.org/10.1177/1948550610377119

[32] Ceren Günsoy and Irmak Olcaysoy Okten. 2022. Inferring Goals and Traits from Behaviors: The Role of Culture, Self-construal, and Thinking Style. *Social Cognition* 40, 2 (2022), 184–211.

[33] Maaike Harbers. 2011. *Explaining agent behavior in virtual training*. Utrecht University.

[34] Maaike Harbers, Joost Broekens, Karel Van Den Bosch, and John Jules Meyer. 2010. Guidelines for developing explainable cognitive models. In *Proceedings of the 10th International Conference on Cognitive Modeling, ICCM 2010*. Drexel University, Philadelphia, PA, 85–90.

[35] Maaike Harbers, Karel Van Den Bosch, and John Jules Meyer. 2010. Design and evaluation of explainable BDI agents. *Proceedings - 2010 IEEE/WIC/ACM International Conference on Intelligent Agent Technology, IAT 2010* 2 (2010), 125–132. https://doi.org/10.1109/WI-IAT.2010.115

[36] Gaole He, Agathe Balayn, Stefan Buijsman, Jie Yang, and Ujwal Gadiraju. 2022. It Is like Finding a Polar Bear in the Savannah! Concept-Level AI Explanations with Analogical Inference from Commonsense Knowledge. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 10, 1 (2022), 89–101. https://doi.org/10.1609/hcomp.v10i1.21990

[37] Gaole He, Lucie Kuiper, and Ujwal Gadiraju. 2023. Knowing About Knowing: An Illusion of Human Competence Can Hinder Appropriate Reliance on AI Systems. *Conference on Human Factors in Computing Systems - Proceedings* (2023). https://doi.org/10.1145/3544548.3581025 arXiv:2301.11333

[38] Fritz Heider. 1958. *The psychology of interpersonal relations*. Wiley;Chapman & Hall.

[39] J. D. Hunter. 2007. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering* 9, 3 (2007), 90–95. https://doi.org/10.1109/MCSE.2007.55

[40] Ramlah Bt. Jantan. 2014. Relationship between Students' Cognitive Style (Field-Dependent and Field-Independent Cognitive Styles) with their Mathematic Achievement in Primary School. *International Journal of Humanities Social Sciences and Education (IJHSSE)* 1, 10 (2014), 2349. www.arcjournals.org

[41] Li Jun Ji, Zhiyong Zhang, and Richard E. Nisbett. 2004. Is it culture or is it language? Examination of language effects in cross-cultural research on categorization. *Journal of Personality and Social Psychology* 87, 1 (2004), 57–65. https://doi.org/10.1037/0022-3514.87.1.57

[42] Frank Kaptein, Joost Broekens, Koen Hindriks, and Mark Neerincx. 2017. Personalised self-explanation by robots: The role of goals versus beliefs in robot-action explanation for children and adults. *RO-MAN 2017 - 26th IEEE International Symposium on Robot and Human Interactive Communication* 2017-Janua, December (2017), 676–682. https://doi.org/10.1109/ROMAN.2017.8172376

[43] Yoshihisa Kashima, Allison McKintyre, and Paul Clifford. 1998. The category of the mind: Folk psychology of belief, desire, and intention. *Asian Journal of Social Psychology* 1, 3 (1998), 289–313. https://doi.org/10.1111/1467-839X.00019

[44] Harold H. Kelley and John L. Michela. 1980. Attribution Theory and Research. *Annual Review of Psychology* 31, 1 (1980), 457–501. https://doi.org/10.1146/annurev.ps.31.020180.002325

[45] Da-jung Kim and Youn-kyung Lim. 2019. Co-Performing Agent: Design for Building User-Agent Partnership in Learning and Adaptive Services. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300714

[46] Jong Hae Kim. 2019. Multicollinearity and misleading statistical results. *Korean journal of anesthesiology* 72, 6 (2019), 558–569.

[47] Khamsum Kinley, Dian Tjondronegoro, Helen Partridge, and Sylvia Edwards. 2012. Human-Computer Interaction: The Impact of Users' Cognitive Styles on Query Reformulation Behaviour during Web Searching. In *Proceedings of the 24th Australian Computer-Human Interaction Conference* (Melbourne, Australia) *(OzCHI '12)*. Association for Computing Machinery, New York, NY, USA, 299–307. https://doi.org/10.1145/2414536.2414586

[48] Nicola Knight and Richard E. Nisbett. 2007. Culture, class and cognition: Evidence from Italy. *Journal of Cognition and Culture* 7, 3-4 (2007), 283–291. https://doi.org/10.1163/156853707X208512

[49] Minkyung Koo, Jongan Choi, and I. Choi. 2018. Analytic versus holistic cognition: Constructs and measurement. In *The Psychological and Cultural Foundations of East Asian Cognition: Contradiction, Change, and Holism*, J. Spencer-Rodgers and K. Peng (Eds.). Oxford University Press, New York, USA, Chapter 4, 105–134. https://doi.org/10.1093/oso/9780199348541.003.0004

[50] Hana Kopecká and Jose Such. 2020. Explainable AI for Cultural Minds. In *European Conference on Artificial Intelligence (Workshop on Dialogue, Explanation and Argumentation for Human-Agent Interaction)*.

[51] Michael W. Kraus, Paul K. Piff, and Dacher Keltner. 2009. Social Class, Sense of Control, and Social Explanation. *Journal of Personality and Social Psychology* 97, 6 (2009), 992–1004. https://doi.org/10.1037/a0016357

[52] Michael W. Kraus, Paul K. Piff, Rodolfo Mendoza-Denton, Michelle L. Rheinschmidt, and Dacher Keltner. 2012. Social class, solipsism, and contextualism: How the rich are different from the poor. *Psychological Review* 119, 3 (2012),

546–572. https://doi.org/10.1037/a0028756
[53] Klaus Krippendorff. 2004. *Content analysis : an introduction to its methodology* (2nd ed. ed.). Sage, Thousand Oaks, Calif.; London.
[54] Todd Kulesza, Simone Stumpf, Margaret Burnett, and Irwin Kwan. 2012. Tell Me More? The Effects of Mental Model Soundness on Personalizing an Intelligent Agent. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Austin, Texas, USA) *(CHI '12)*. Association for Computing Machinery, New York, NY, USA, 1–10. https://doi.org/10.1145/2207676.2207678
[55] Lee Dong Kyu Lee Sangseok. 2018. What is the proper way to apply the multiple comparison test? *kja* 71, 5 (2018), 353–360. https://doi.org/10.4097/kja.d.18.00242 arXiv:http://www.e-sciencecentral.org/articles/?scid=1156583
[56] Tania Lombrozo. 2007. Simplicity and probability in causal explanation. *Cognitive Psychology* 55, 3 (2007), 232–257. https://doi.org/10.1016/j.cogpsych.2006.09.006
[57] Bertram F Malle. 1999. How People Explain Behavior: A New Theoretical Framework. *Personality and Social Psychology Review* 3, 1 (1999), 23–48. https://doi.org/10.1207/s15327957pspr0301_2
[58] Bertram F Malle. 2011. Attribution Theories: How People Make Sense of Behavior. *Theories in Social Psychology*. 23 (2011), 72–95.
[59] Bertram F. Malle. 2011. Time to give up the dogmas of attribution. An alternative theory of behavior explanation. In *Advances in experimental social psychology*. ADVANCES IN EXPERIMENTAL SOCIAL PSYCHOLOGY, Vol. 44. Elsevier, San Diego, 297–352.
[60] Bertram F. Malle. 2014. *F.EX: A Coding Scheme for Folk Explanations of Behavior*. Brown University.
[61] Bertram F. Malle, Joshua M. Knobe, and Sarah E. Nelson. 2007. Actor-Observer Asymmetries in Explanations of Behavior: New Answers to an Old Question. *Journal of Personality and Social Psychology* 93, 4 (2007), 491–514. https://doi.org/10.1037/0022-3514.93.4.491
[62] Hazel Markus and Shinobu Kitayama. 1991. Culture and the self: Implications for cognition, emotion, and motivation. American Psychological Association. *Psychological review* 98, 2 (1991), 224–253. http://web.b.ebscohost.com.libezproxy.open.ac.uk/ehost/pdfviewer/pdfviewer?vid=1&sid=620938ca-e7f2-44a8-8575-fff36c02dead%40pdc-v-sessmgr02
[63] Charles Martin, Henry Gardner, Ben Swift, and Michael Martin. 2016. Intelligent Agents and Networked Buttons Improve Free-Improvised Ensemble Music-Making on Touch-Screens. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 2295–2306. https://doi.org/10.1145/2858036.2858269
[64] Takahiko Masuda and Richard E. Nisbett. 2001. PERSONALITY PROCESSES AND INDIVIDUAL DIFFERENCES Attending Holistically Versus Analytically: Comparing the Context Sensitivity of Japanese and Americans. *Journal of Personality and Social Psychology* 81, 5 (2001), 922–9343. https://doi.org/10.1037/0022-35I4.81.5.922
[65] Scott Mayer McKinney, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back, Mary Chesus, Greg S Corrado, Ara Darzi, et al. 2020. International evaluation of an AI system for breast cancer screening. *Nature* 577, 7788 (2020), 89–94.
[66] Mehmet Mehmetoglu. 2018. Medsem: A Stata package for statistical mediation analysis. *International Journal of Computational Economics and Econometrics* 8, 1 (2018), 63–78.
[67] Christian Meske, Enrico Bunde, Johannes Schneider, and Martin Gersch. 2022. Explainable Artificial Intelligence: Objectives, Stakeholders, and Future Research Opportunities. *Information Systems Management* 39, 1 (2022), 53–63. https://doi.org/10.1080/10580530.2020.1849465
[68] Martijn Millecamp, Robin Haveneers, and Katrien Verbert. 2020. Cogito ergo quid? the Effect of Cognitive Style in a Transparent Mobile Music Recommender System. In *UMAP 2020 - Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*. Association for Computing Machinery, New York, NY, USA, 323–327. https://doi.org/10.1145/3340631.3394871
[69] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence* 267 (2019), 1–38.
[70] Tim Miller, Piers Howe, and Liz Sonenberg. 2017. Explainable AI: Beware of inmates running the asylum or: How I learnt to stop worrying and love the social and behavioural sciences. *arXiv preprint arXiv:1712.00547* (2017).
[71] Sina Mohseni, Niloofar Zarei, and Eric D. Ragan. 2021. A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. *ACM Trans. Interact. Intell. Syst.* 11, 3–4, Article 24 (aug 2021), 45 pages. https://doi.org/10.1145/3387166
[72] Jonathan Morgan, Connor Wood, and Catherine Caldwell-Harris. 2018. Reflective thought, religious belief, and the social foundations hypothesis. In *The New Reflectionism in Cognitive Psychology: Why Reason Matters*, Gordon Pennycook (Ed.). Routledge, Abingdon, Chapter 2, 10–32. https://doi.org/10.4324/9781315460178
[73] Michael W. Morris and Kaiping Peng. 1994. Culture and Cause: American and Chinese Attributions for Social and Physical Events. *Journal of Personality and Social Psychology* 67, 6 (1994), 949–971. https://doi.org/10.1037/0022-3514.67.6.949

[74] Francesca Mosca and Jose Such. 2022. An explainable assistant for multiuser privacy. *Autonomous Agents and Multi-Agent Systems (JAAMAS)* 36, 10 (2022), 1–45.
[75] NHS. [n. d.]. *Diabetes*. https://www.nhs.uk/conditions/diabetes/
[76] Richard E. Nisbett, Kaiping Peng, Incheol Choi, and Ara Norenzayan. 2001. Culture and systems of thought: Holistic versus analytic cognition. *Psychological Review* 108, 2 (2001), 291–310. https://doi.org/10.1037//0033-295x.108.2.291
[77] Edward C. Norton, Bryan E. Dowd, and Matthew L. Maciejewski. 2019. Marginal Effects—Quantifying the Effect of Changes in Risk Factors in Logistic Regression Models. *JAMA* 321, 13 (04 2019), 1304–1305. https://doi.org/10.1001/jama.2019.1954
[78] Mahsan Nourani, Joanie King, and Eric Ragan. 2020. The Role of Domain Expertise in User Trust and the Impact of First Impressions with Intelligent Systems. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 8 (2020), 112–121. https://doi.org/10.1609/hcomp.v8i1.7469 arXiv:2008.09100
[79] Irmak Olcaysoy Okten and Gordon B. Moskowitz. 2020. Spontaneous goal versus spontaneous trait inferences: How ideology shapes attributions and explanations. *European Journal of Social Psychology* 50, 1 (2020), 177–188. https://doi.org/10.1002/ejsp.2611
[80] Bruno Uchenna Onyekuru. 2015. Field Dependence-Field Independence Cognitive Style, Gender, Career Choice and Academic Achievement of Secondary School Students in Emohua Local Government Area of Rivers State. *Journal of Education and Practice* 6, 10 (2015), 76–85.
[81] Nelly Oudshoorn, Els Rommes, and Marcelle Stienstra. 2004. Configuring the user as everybody: Gender and design cultures in information and communication technologies. *Science, Technology, & Human Values* 29, 1 (2004), 30–63.
[82] Daphna Oyserman, Heather M. Coon, and Markus Kemmelmeier. 2002. Rethinking individualism and collectivism: Evaluation of theoretical assumptions and meta-analyses. *Psychological Bulletin* 128, 1 (2002), 3–72. https://doi.org/10.1037/0033-2909.128.1.3
[83] Ruth A. Palmquist and Kyung-Sun Kim. 2000. Cognitive style and on-line database search experience as predictors of Web search performance. *Journal of the American Society for Information Science* 51, 6 (2000), 558–566. https://doi.org/10.1002/(SICI)1097-4571(2000)51:6<558::AID-ASI7>3.0.CO;2-9 arXiv:https://asistdl.onlinelibrary.wiley.com/doi/pdf/10.1002/%28SICI%291097-4571%282000%2951%3A6%3C558%3A%3AAID-ASI7%3E3.0.CO%3B2-9
[84] Leslie E. Papke and Jeffrey M. Wooldridge. 1996. Econometric methods for fractional response variables with an application to 401(k) plan participation rates. *Journal of Applied Econometrics* 11, 6 (1996), 619–632. https://doi.org/10.1002/(SICI)1099-1255(199611)11:6<619::AID-JAE418>3.0.CO;2-1
[85] Vikram V. Ramaswamy, Sunnie S. Y. Kim, Ruth Fong, and Olga Russakovsky. 2022. Overlooked factors in concept-based explanations: Dataset choice, concept learnability, and human capability. (2022). arXiv:2207.09615 http://arxiv.org/abs/2207.09615
[86] Anand Rao and Michael Georgeff. 1995. BDI Agents: From Theory to Practice. *Proceedings of the First International Conference on Multi-Agent Systems (ICMAS-95)* 95 (1995), 312–319.
[87] Natalia Reich and Friederike Eyssel. 2015. Attitudes towards service robots in domestic environments: The role of personality characteristics, individual interests, and demographic variables. *Paladyn, Journal of Behavioral Robotics* 4, 2 (2015), 123–130. https://doi.org/10.2478/pjbr-2013-0014
[88] Mireia Ribera and Àgata Lapedriza. 2019. Can we do better explanations? A proposal of user-centered explainable AI. In *Joint Proceedings of the ACM IUI 2019 Workshops*. ACM, New York, NY, USA, 7 pages.
[89] Vincent Robbemond, Oana Inel, and Ujwal Gadiraju. 2022. Understanding the Role of Explanation Modality in AI-assisted Decision-making. *UMAP2022 - Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization* (2022), 223–233. https://doi.org/10.1145/3503252.3531311
[90] Amir Saemi. 2015. Aiming at the good. *Canadian Journal of Philosophy* 45, 2 (2015), 197–219. https://doi.org/10.1080/00455091.2015.1054230
[91] Massimiliano Scopelliti, Maria Vittoria Giuliani, and Ferdinando Fornara. 2005. Robots in a domestic setting: a psychological approach. *Universal access in the information society* 4, 2 (2005), 146–155.
[92] Amitai Shenhav, David G. Rand, and Joshua D. Greene. 2012. Divine intuition: Cognitive style influences belief in God. *Journal of Experimental Psychology: General* 141, 3 (2012), 423–428. https://doi.org/10.1037/a0025391
[93] Jacob Sohlberg. 2015. Thinking Matters: The Validity and Political Relevance of Need for Cognition. *International Journal of Public Opinion Research* 28, 3 (08 2015), 428–439. https://doi.org/10.1093/ijpor/edv023 arXiv:https://academic.oup.com/ijpor/article-pdf/28/3/428/6771546/edv023.pdf
[94] Ben Steichen and Bo Fu. 2020. Cognitive Style and Information Visualization—Modeling Users Through Eye Gaze Data. *Frontiers in Computer Science* 2 (2020). https://doi.org/10.3389/fcomp.2020.562290
[95] Jose Such. 2017. Privacy and autonomous systems. *IJCAI International Joint Conference on Artificial Intelligence* 0 (2017), 4761–4767. https://doi.org/10.24963/ijcai.2017/663

[96] Thomas Talhelm, Jonathan Haidt, Shigehiro Oishi, Xuemin Zhang, Felicity F. Miao, and Shimin Chen. 2015. Liberals Think More Analytically (More "WEIRD") Than Conservatives. *Personality and Social Psychology Bulletin* 41, 2 (2015), 250–267. https://doi.org/10.1177/0146167214563672

[97] S. Talhelm, Thomas, Zhang, X., Oishi, S., Shimin, C., Duan, D., Lan, X., Kitayama. 2014. Large-Scale Psychological Rice Versus Wheat Agriculture. *Science* 344, May (2014), 603–608. arXiv:arXiv:1011.1669v3

[98] Sam Thellman, Maartje de Graaf, and Tom Ziemke. 2022. Mental State Attribution to Robots: A Systematic Review of Conceptions, Methods, and Findings. *ACM Transactions on Human-Robot Interaction* 11, 4 (2022), 1–51. https://doi.org/10.1145/3526112

[99] Tom van Nuenen, Xavier Ferrer, Jose Such, and Mark Cote. 2020. Transparency for Whom? Assessing Discriminatory Artificial Intelligence. *IEEE Computer* 53 (2020), 36–44.

[100] Tom van Nuenen, Jose Such, and Mark Coté. 2022. Intersectional Experiences of Unfair Treatment Caused by Automated Computational Systems. In *Proceedings of the ACM on Human-Computer Interaction-CSCW*. ACM.

[101] Tom van Nuenen, Jose Such, and Mark Coté. 2022. Intersectional Experiences of Unfair Treatment Caused by Automated Computational Systems. In *PACM on Human-Computer Interaction - ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW)*, Vol. 6. 1–30.

[102] Michael E.W. Varnum, Igor Grossmann, Shinobu Kitayama, and Richard E. Nisbett. 2010. The origin of cultural differences in cognition: The social orientation hypothesis. *Current Directions in Psychological Science* 19, 1 (2010), 9–13. https://doi.org/10.1177/0963721409359301

[103] Rachael C Walker, Roger Marshall, Kirsten Howard, Rachael L Morton, and Mark R Marshall. 2017. "Who matters most?": Clinician perspectives of influence and recommendation on home dialysis uptake. *Nephrology* 22, 12 (2017), 977–984.

[104] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y. Lim. 2019. Designing Theory-Driven User-Centric Explainable AI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3290605.3300831

[105] Sarah Myers West, Meredith Whittaker, and Kate Crawford. 2019. Discriminating systems.

[106] Michael Winikoff, Galina Sidorenko, Virginia Dignum, and Frank Dignum. 2021. Why bad coffee? Explaining BDI agent behaviour with valuings. *Artificial Intelligence* 300 (2021), 103554. https://doi.org/10.1016/j.artint.2021.103554

[107] H. A. Witkin, C. A. Moore, D. Goodenough, and P. W. Cox. 1977. Field-Dependent and Field-Independent Cognitive Styles and Their Educational Implications. *Review of Educational Research* 47, 1 (1977), 1–64. https://doi.org/10.3102/00346543047001001

[108] Michael Wooldridge. 2003. *Introduction to MultiAgent Systems*. Wiley, Hoboken.

[109] Michael Wooldridge and Nicholas R. Jennings. 1995. Intelligent agents: theory and practice. *The Knowledge Engineering Review* 10, 2 (1995), 115–152. https://doi.org/10.1017/S0269888900008122

[110] Xinshu Zhao, John G Lynch Jr, and Qimei Chen. 2010. Reconsidering Baron and Kenny: Myths and truths about mediation analysis. *Journal of consumer research* 37, 2 (2010), 197–206.

Received January 2023; revised July 2023; accepted November 2023