On the importance of experimental psychology for explainable artificial intelligence [*]

Wai Wong[0000−0002−5442−2562], Walter Schaeken[0000−0002−5220−1643], and Joost Vennekens[0000−0002−0791−0176]

KU Leuven, 3000 Leuven, Belgium
{wai.wong,walter.schaeken,joost.vennekens}@kuleuven.be

Abstract. Explainable AI (XAI) has gained popularity in research in recent years since we human users like to understand why AI arrives at a particular decision and behaves as it does. However, a recent paper written by Miller et al. [19] has shown that most XAI models are not built on the current scientific understanding of human explanation, let alone tested with human behavioural experiments. In this position paper, we argue why experimental methods derived from psychology are crucial in advancing XAI research. In addition, by focusing on theories in folk psychology, we can see what is left to be done for us to equip XAI models with commonsense reasoning. Moreover, by looking back at the successful applications of experimental psychology in engineering and cognitive science, from Human Factors Engineering (HFE), and Human-Computer Interaction (HCI) to Experimental Pragmatics, insights on the collaboration with psychologists can be drawn in to pave the way to take XAI research to a whole new level.

Keywords: Explainable artificial intelligence (XAI) · Experimental psychology · Folk psychology · Commonsense reasoning

1 Introduction: Social Insights in Explanation

Explainable artificial intelligence (XAI) has recently experienced a surge of attention as researchers and practitioners strive to make their algorithms more transparent. Miller's influential paper [18] directed the attentions of XAI research to social insights in explanation. Since much XAI research focuses on explicitly explaining decisions or actions to a human observer, it should not be controversial to suggest that observing how humans explain things to one another can be a useful starting point for building XAI. However, it is fair to say that the vast majority of work in this area is based solely on the researchers' assumptions about what constitutes a "good" explanation. Extensive and valuable research in philosophy, psychology, and cognitive science on how people define,

[*] The first author's work is currently supported by a doctoral grant from the Internal Funds KU Leuven (BIT-SHARE-project IDN/19/009). The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

generate, select, evaluate, and present explanations contend that cognitive biases and social expectations are used during the explanation process. Miller's paper [18] argues that the field of explainable artificial intelligence can build on existing research and reviews relevant papers from philosophy, cognitive science, and social psychology. Indeed, as Miller and his colleagues pointed out in 2017 [19], the majority of XAI models are not based on current scientific understanding of human explanation, let alone tested with human behavioural experiments. Miller [18] concluded that the following four social insights should be considered in XAI models:

- why-questions are contrastive;
- explanations are chosen (biasedly);
- explanations are social; and
- probabilities do not matter as much as causal links.

In recent years, social insights in explanation research have evolved at a breakneck pace. As an example, Kirfel et al. [12] investigated how humans communicate with causal explanations in order to clarify the distinct roles that normalcy and causal structure play in causal judgement and pave the way for a more comprehensive explanation of the causal explanation. They provide evidence that causal explanations routinely reveal much more than this fundamental information, in addition to providing a communication-theoretical account of explanation that makes precise predictions about the kinds of inferences people will make from other people's explanations. However, our position paper does not aim to provide an update on research progress since Miller's paper was published [18]. On the contrary, we will dig deeper into the social aspect of explanation and argue why experimental methods derived from psychology are critical for advancing XAI research.

One thing we want to be upfront about regarding this position paper: we are not suggesting that current XAI researchers should abandon their approaches. On the contrary, the advancement of XAI models is critical to the overall progress made in this field. Without these models, it is impossible for the field to progress. This position paper simply emphasises that certain challenges and problems in XAI research cannot be solved solely by developing AI models. By bringing this issue to light, we hope that all XAI researchers will recognise the limitations of their model-building focus and, whenever possible, invite experimental psychologists to participate in model testing.

Miller's paper [18] established that human explanation and reasoning are inherently social, but what exactly is "explanation"? Lombrozo [14] defines explanation as a process as well as a product. Meanwhile, Miller [18] contends that there are two processes in play in addition to the product:

- Inferential process [1] — It is an abductive inference process used to fill a gap in an explanation for a specific event. In social science, the process of

[1] The inferential process is originally named as the cognitive process by Miller [18]. To avoid confusion with cognitive models introduced in the later section of this paper, it is renamed as the inferential process.

determining the causes of a specific phenomenon is known as attribution, and it is only one step in the overall explanation process.
- Product — The product of the inferential explanation process is the explanation that results from the process.
- Social process — The goal of the social process, which typically involves group interactions, is to transfer knowledge between the explainer and the explainee so that the explainee understands the causes of the event.

Given the emphasis on the social aspect in the explanation, it should come as no surprise that human-agent interaction, the field in which XAI research is located, is actually a convergence of social science (e.g., psychology), artificial intelligence, and human-computer interaction [18]. In the remainder of the paper, we will first introduce related folk psychological theory on explanation (sections 2 and 3), and then we will discuss the historical relationship between psychology and engineering (section 4). By discussing its contribution to human-computer interaction research (section 5), we discuss the effort in formalising folk psychological theory for computational use (section 6), as well as how psychology has successfully transformed its neighbouring fields (section 7).

2 Cognitive models vs Inferential theory

The main goal of this paper is to encourage XAI researchers to collaborate with experimental psychologists. In fact, our call to action is not the first of its kind. Taylor and Taylor [27] made a similar call for the involvement of experimental psychologists in XAI research in their article “Artificial cognition: How experimental psychology can help generate explainable artificial intelligence” in 2020. As Taylor and Taylor [27] clearly demonstrate, deep neural network-based artificial intelligence has developed to the point where it can be challenging or impossible to explain how a model comes to its conclusions. This black-box issue is particularly troubling when the model makes judgments that have a potential impact on people’s well-being. This is also what motivates the XAI research, which tries to improve machine learning’s interpretability, fairness, and transparency. Taylor and Taylor [27] argue that the experience of cognitive psychology dealing with the mind’s black box by means of experimental research makes it clear that cognitive psychology can contribute to the development of XAI. They argue that in order to increase explainability, the principles and standards of experimental cognitive psychology should be applied when investigating artificial black boxes. One of the pillars of experimental cognitive psychology is that the goal of an experiment is to falsify the null hypothesis or competing hypotheses. With this process, theories can be developed and refined. So looking for reasonable counter-explanations for models is crucial. Furthermore, using standardised tasks, looking carefully for variations in the outcome and finding boundary conditions are other examples of their many recommendations inspired by experimental cognitive psychology. Although extremely useful and relevant, Taylor and Taylor’s paper [27] focuses on one of the two modes of improving AI explainability — by making the content of the AI cognitive models more interpretable. The other — equipping AI models with inferential theory — will be the focus of this paper instead.

What is the difference between inferential theory and cognitive models? According to Gordon and Hobb's book [8], computational cognitive models and inferential theories are the end results of two very different lines of research. The book provides examples of how cognitive models and inferential theory differ from one another. On the one hand, cognitive models are developed *by **cognitive scientist*** to advance the study of human cognition, i.e. *the study of how people — as cognitive agents — think*. In the case of AI, for example, cognitive models are used to support the development of robotics. On the other hand, inferential theories are developed — *mostly by **cognitive agents** themselves — to explain how they think about something*, the knowledge that underpins the cognitive processes of explanation and prediction. The analogy in AI for inferential theories would be the knowledge representation research. This distinction can be difficult to understand when inferential theories are concerned with cognitive processes. A specific computational cognitive model may share many conceptual similarities with inferential theories of folk (commonsense) psychology. The real difference between these two classes of models is in the methods used to evaluate them. For example, a cognitive model of human emotions and an inferential theory of human emotions may address the same mental states and processes and may even use some of the same terminology, logic, and constructs. A cognitive model of human emotions will be evaluated based on how well it replicates empirical data on emotional behaviour in people. The effectiveness of an inferential theory of human emotions is determined by how well it simulates the hypotheses and justifications that people come up with when thinking about human emotion behaviour.

Typically, Marr's level of analysis [16] is used to study cognitive models, which has three levels and each has its own question(s):

- Computation — what issues does it address or resolve?
- Representation & Algorithm — how does the system carry out its functions? Specifically, how does it construct and manipulate representations, and what representations does it use?
- Implementation — how is the system actually implemented, i.e. what is the process of moving from abstract thought to actualized behaviour?

In practice, because most computational languages, even declarative ones, are mono-inferential, or based on only one type of inference, representation and algorithm are typically viewed as complementary. Due to the mono-inferential features of computational languages, it is difficult to reuse representations of the same information because different inference engines typically have their own specialised computational languages. The Knowledge Base (KB) paradigm proposed by Denecker and Vennekens [4] is motivated by this concern about knowledge reusability. The KB paradigm strictly separates informational concerns (such as knowledge representation) from problem-solving concerns (i.e. from goals to plans and execution). A knowledge base system provides various inference techniques and allows information to be stored in a knowledge base.

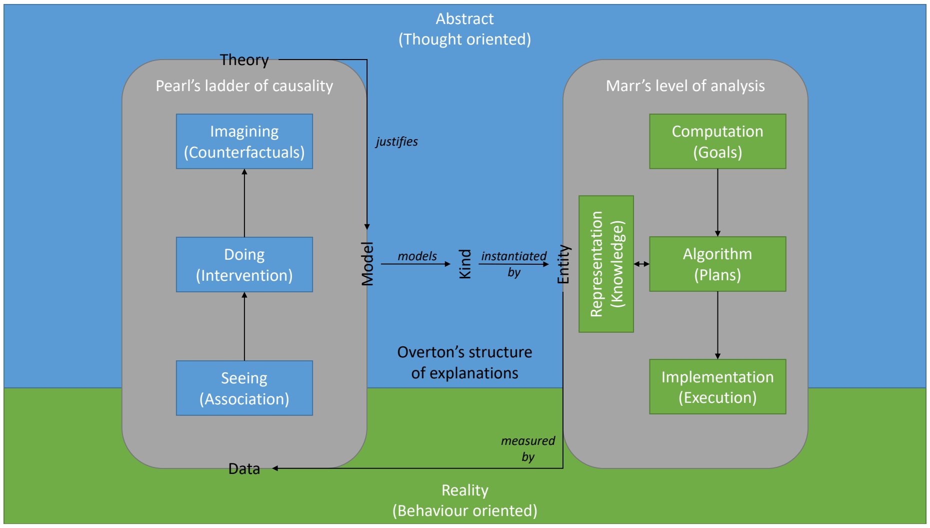

Fig. 1. Marr's level of analysis versus Pearl's ladder of causality, connected with Overton's structure

These inference techniques allow the same knowledge base to be used to solve a variety of tasks and problems. By itself, the knowledge base cannot be run or executed because it is neither a programme nor a description of a problem. It is just information. However, this information can be used to solve a variety of problems. As a result, the KB paradigm is multi-inferential [29].

In contrast to Marr's level of analysis [16], which examines a system's behaviour based on its internal processes, Pearl's ladder of causality [22] proposes various modelling rungs from data to theory:

- from "Seeing" — i.e. one object is associated with another if the probability of observing one changes the probability of observing the other
- and "Doing" — i.e. this level asserts specific causal relationships between events)
- to finally "Imagining" — i.e. the highest level, counterfactual, involves consideration of an alternate version of a past event, or what would happen under different circumstances for the same experimental unit.

We hypothesise that the ability to imagine helps humans develop inferential theories of how people think. Using Overton's explanation framework [21] mentioned in Miller's paper [18], we also propose the following link between Marr's level and Pearl's ladder (see figure 1). "Model" — which can refer to any rung of Pearl's ladder — models "Kind" which is instantiated by "Entity" with Marr's level. In abstract thought, "Theory" — at the top of Pearl's ladder — can be used to support modelling at each rung, whereas "Data" — at the bottom of Pearl's ladder — is a measurement of the "Entity" in reality. Moving up Pearl's ladder from one rung to the next requires explicit thought. In contrast, once various Marr's levels are specified, the analysis in each level, i.e. computation, algorithm and implementation, can be carried out automatically. This is consistent with psychology's dual process theory [6], where we can draw the analogy between Pearl's ladder and System 2 — an explicit, controlled, and conscious process — and between Marr's levels and System 1 — an implicit, automatic and unconscious process. Human inferential theory falls under System 2 because it explicitly models and speculates on the mental states of others. In the following section, we will go over how inferential theory leads to human explanation.

Before proceeding to the next section, we should note that a recent Robotics Reasoning Architecture by Sridharan and Meadows [25] also exhibits this dual process nature. The architecture is envisioned as a collaboration between a statistician (focus on behaviour with automatic nature, i.e. System 1) and a logician (focus on thought with controlled nature, i.e. System 2), combining the complementary strengths of declarative programming, probabilistic reasoning, and relational learning. It represents and explains the world and the robot's understanding of it at two granularities. A fine-resolution description of the domain is reasoned about using non-monotonic logic and is close to the data obtained from the robot's sensors and actuators, whereas a coarse-resolution description of the domain includes common sense knowledge. Despite the fact that Sridharan and Meadows' architecture does not use a single unified logical-probabilistic representation, it establishes and precisely defines a tight coupling between the representations at the two granularities, allowing the robot to represent and reason about commonsense knowledge, and what it knows (or does not know), and how actions affect the robot's knowledge. A conversation between a logician and a statistician, as well as their physical and mental actions, are interpreted as exemplifying how the two types of knowledge interact with the corresponding reasoning techniques.

3 Folk Psychology: Turing test & Malle's theory of human explanation

In the previous section, we have introduced the concept of inferential theory which describes the way that people think about something. What if that something is the way how one thinks? Then we enter the realm of folk psychology (aka commonsense psychology), which is the study of the human capacity to explain and predict the behaviour and mental state of other people. Interestingly, we use this capacity not only on humans but sometimes on inanimate things as well, including machines, to act like machines could think. We will discuss this anthropomorphic use further in section 6. Nevertheless, we can still tell this machine thinking is just a pretence created by our own thought. But what if machines are intelligent enough and act in such a way that we can no longer tell the difference?

In fact, this machine intelligence question inspired Alan Turing's article "Computing Machinery and Intelligence," which launched the field of artificial intelligence in 1950 [28]. Turing devised the "Turing Test," a hypothetical test in which he first defined intelligence as a human judgement before describing the test itself: Is it possible for a computer to fool someone into thinking it is a real person? Gorden and Hobb rephrase this question by removing the aspect of misdirection: "Does this computer operate the way people think they think?" [8].

All folk explanations of others' behaviour use the implicit theories about their beliefs, objectives, plans, and emotions. This kind of explanation is also known as social attribution. How do we make social attribution? Miller [18] asserts that Malle's model [15] is the most developed and comprehensive social attribution model to date. Traditional explanations for human behaviour, in various forms, are based on Malle's concept of intentionality and its essential elements of belief, desire, and intention. In general, for an action to be deemed intentional, all five elements of the intentionality concept must be present: the agent's desire for the outcome, belief that the action would produce the desired outcome, intention to carry out the action, skill to carry out the action, and awareness of achieving the intention while carrying out the action. *Reason explanations* are the main method used to explain intentional behaviours. They contain the justifications for an agent's intentions or deliberate actions. Reasons include background beliefs or desires that are informative and that relate to the desire and belief elements of intentionality. Contrarily, unintentional behaviours are not the result of intentions or belief-desire reasoning; rather, they can be attributed to a wide range of factors, including physiology and culture [15].

It is common practice for us to anthropomorphize machines as if they have intentions, which is not surprising given that we do the same thing with animals and even inanimate objects. In fact, we have programmed the machine in such a way that its imperative procedures reflect the designer's or user's intention, which is frequently — and perhaps incorrectly — referred to as the machine's intention. In other words, we have made ourselves think that machine has its own intention. Since the explanation given by XAI are intended to assist humans in better understanding the actions of AI, humans may hold AI explanations to the same standard as human explanations. What kinds of things do humans consider to be reasonable? What does the human consider to be a valid and objective explanation? These questions are what we believe pushes XAI research from the realm of computer science to the realm of psychology, as psychology has been studying the human mind — including human judgement — for decades. In the next section, we start looking at the history of psychology and identify the root of the collaboration between psychology and computer science.

4 Engineering roots in Psychology

What is psychology? Psychology is the study of the human mind and human behaviour from their physiological basis (biopsychology), the process of information (cognitive psychology), individual characteristics (psychology of individual differences), response to social settings (social psychology) and well-being (psychopathology and health psychology) towards how they develop in life (developmental psychology) and evolve through history (evolutionary psychology). Although psychology has a long history (already the Greek philosophers were asking psychological questions), it is only in the nineteenth century that psychology became an independent science in which experimental data collection was crucial. In Germany, Gustav Fechner theorized in 1954 about judgments of sensory experiences and later conducted some experiments. Some year later, Wilhelm Wundt founded the first Psychology Laboratory. Around the same time, William James started a small experimental psychology laboratory, which was, however, mainly used for demonstration purposes [3].

One of the motives for developing these laboratories can be traced back to engineering, and more precisely a practical concern with optical technology (see Wilson et al., 2012) [31]. Indeed, in the beginning of the 19th century, it was recognized that observations by astronomers are not that straightforward and are influenced by individual differences. This interest started after a clash in the Royal Observatory in Greenwich between two astronomers. The assistant astronomer Kinnebrook reported systematically different times at which a star crossed the marker in a telescope than the Royal Astronomer Maskelyne did. This clash inspired research that led to the discovery of the personal equation: there is an inherent bias in individual measurements and observations [23]. The personal equation turned out not only to be important for astronomy, but also for other commercial and military operations, and became the topic of investigation in the labs of Wundt and James [31].

Given the engineering roots of psychology, it is no surprise that engineering psychology (aka human factors engineering) — the science of human behaviour and capability, applied to the design and operation of systems and technology, e.g. human-machine interaction — originated from within experimental psychology soon after psychology emerges as an academic discipline. The two world wars played an important role in this field. Failing weapons and wrongly dropped bombs had not only technical causes but also human errors were causing them. Experimental psychologists helped not only to understand these mistakes but also to find solutions. For instance, the noise levels in military aircraft turned out to be one of the causes leading to human errors and psychologists Stevens and Beranek developed improved microphones and earphones for communication in airplanes. Such collaborations between different fields of science (natural science, engineering, psychology) grew after World War II and improved man-machine interactions considerably, but also enhanced the development of theories in psychology itself. Signal detection theory for instance is a nice example of these efforts [5].

To summarise, the birth of psychology can be traced back to solving an engineering problem: the interaction between humans and technology. In the next section, we will focus on human interaction with a specific type of machine — computers.

5 Interaction with machines: From operations to communication

Human-computer interaction (HCI) is a field of study that focuses on the interfaces that humans use to interact with computers. Researchers in HCI study how people use computers and develop technologies that let people use them in fresh ways. Human-computer interaction is a research area that straddles a number of academic disciplines, including computer science, psychology, design, media studies, and behavioural sciences. The phrase HCI is popularised In Card et al. 's 1983 book "The Psychology of Human-Computer Interaction" [2]. The term is meant to convey that, in contrast to other tools with specific and constrained uses via operations, computers have a wide range of uses, many of which involve an ongoing conversation between the user and the computer via communications. The idea of dialogue compares human-computer interaction to interpersonal communication, which is an analogy that is essential to the field's theoretical foundations, as well as the shift from the operation of machines to the communication with them.

Carbonnelle et al's work [1] on interactive consultants is one instance with a focus on communication with machines. In their interaction, the user and the system jointly construct a model of a given knowledge base, i.e. a situation that satisfies all of the constraints in the knowledge base. Additionally, the user can ask the interactive consultant to explain how it comes its conclusions. It then lists the user-provided pertinent data about the specific circumstances as well as the laws from the knowledge base that led to the derivation. According to various psychological research [20][13][17], people reason more effectively when collaborating with others as they receive immediate feedback in this setting. We speculate such benefits will also be seen in human-machine collaboration if the machine can provide immediate feedback, like Carbonnelle's interactive consultants.

There is no doubt that there is a better way than just following the human-machine interaction tradition to create human-computer interfaces, according to many researchers, experts and practitioners in the computer field. Sadly, they cannot agree on what it is. Some people consider it obvious that psychological knowledge should be used. In the words of Hansen, [10], their motto could be "Know the user!". But when it comes to talking about how humans interact with technology, it is frequently assumed that all one needs are ways to make sure that the obvious is not missed; "All we need from psychology is a few good checklists!" might be the catchphrase in this case. However, Card et al. assert in their book (1983) that checklists cannot capture all aspects of human-computer interaction [2].

Engineering psychology suggests that psychology should be involved in the design of the user-computer interface. It is because the success of psychology in improving the flyability of airplanes leads us to believe that improving the usability of computers through the same psychological attention to human performance is possible. Engineering psychology, or human factors engineering, has excelled at evaluation. With a real system, one can produce a judgement via an experiment. The methodology of experimental design, supported by concurrent expertise in experimental control and statistics to assess the results, has thus been the main tool in the human-factors toolbox. The emphasis on evaluation is widespread, e.g. evaluating social action programs is the focus of a whole subfield of psychology. Whether it is concerned with clinical evaluation or intelligence testing, the testing movement is fundamentally evaluative in nature [2].

Following the insights gained from prior work, Sridharan & Meadows [26] have identified the following guiding principles or claims to support explanations in human-robot collaboration:

1. Appropriate — At a suitable level of abstraction, explanations should present context-specific information pertinent to the domain, task, or question at hand;
2. On-demand transparency — Online descriptions of decisions, justifications for decisions, knowledge, beliefs, experiences that shaped the beliefs, and underlying strategies or models should all be available in explanations;
3. General-purpose components — There should be as little task- or domain-specific content in explanation generation systems as possible;
4. Human in the loop — Systems that generate explanations should take into account human comprehension and feedback to guide their decisions.
5. Non-monotonic reasoning — Systems for generating explanations should make use of knowledge components that allow for non-monotonic revision based on observations made right away or later, as a result of active exploration or the execution of reactive actions.

Concepts behind most of the aforementioned guiding principles are discussed in one way or the other throughout our paper. (1) We can see the first principle is similar to the cooperation principle when human agents are involved in the conversation, i.e. following the four goals: be truthful, say just enough, stay relevant and be clear. These cooperation principles, also known as Gricean Maxims [9] in pragmatics — a linguistics subject that studies the use of language and how context contributes to meaning. In section 7, we will explore how experimental psychology transforms the study of pragmatics. (2) Meanwhile, the on-demand transparency principle is following the motivation behind the surge of XAI research and hopes to mitigate the lack of transparency and interpretability issues in current popular black-box models e.g. artificial neural network (ANN) and deep learning. (3) The general-purpose components principle is addressed by the knowledge-based paradigm discussed in section 2, as it aims to move beyond the mono-inferential knowledge representation in declarative languages. (4) Human-in-the-loop is an important principle, which we will discuss further with a focus on the accommodation of human anthropomorphic tendencies in the next section. (5) Finally, humans mostly rely on non-monotonic reasoning, but this discussion is outside the scope of the current paper.

The takeaway message from the development of the HCI study: only having a checklist of social insights are not enough, we also need to keep human in the loop! To do so, we need human evaluation and this can be done via enlisting the help pf psychologists, which is further discussed in section 7.

6 Difficulties in anthropomorphic computing: the lack of psychological understanding

In section 3, we briefly mentioned the human tendency to anthropomorphise objects as if they have intentions even though we know they do not. Nevertheless, personal computers today are still created using a totally different metaphor. Despite the points of attention on HCI that were already written down almost four decades ago [2], Gordan & Hobbs [8] have argued that most developers to date still want users to treat computers like office space rather than as living beings. They list out three major sources of discordance we are still facing in human-computer interaction:

- The first major source of conflict is that computers are unable to meet or even acknowledge our commonsense psychological expectations.
- The poor understanding that computers have of their users is a second source of disagreement. Though personalization and user modelling are increasingly common in modern computing systems, few make an effort to take user behaviour into account outside of what is immediately observable.
- The third point of discordance, which is possibly the most troubling, is that computers have little to no ability to comprehend human language related to psychology (e.g. thinking, feelings, emotions), despite the fact that it is crucial to human-human communication.

We suspect most of the current Human-AI interactions still follow similar discordances.

To respond to these discordances, an engineering strategy called anthropomorphic computing is devised. The main tenet of this strategy is that computational systems should support their users' anthropomorphic tendencies by acting in a way that fits their realistic psychological models. Computational systems will inevitably be regarded by humans as having intentions, objectives, beliefs, expectations, and emotions. Instead of using engineering models that are used in the architectural design of computing hardware and software, AI will logically predict and explain the reasoning behaviour of computational systems using the same commonsense psychological models that are used by humans to predict and explain human reasoning behaviour. Human users will be able to apply their commonsense psychological models to their computers with verisimilitude if computational systems' reasoning behaviour is aligned with common sense psychological models. [8]

One popular counterargument against anthropomorphic computing is that computers would have to do their computation following the human reasoning style, which is known as full of irrational bias and less optimal. Like airplanes do not need to fly like birds, computers do not have to think like a human. But in order to communicate efficiently with humans, they need to know how humans think. We, humans, understand how computers think, but we do not think like a computer. We can act like it, but most of the time we do not think like a computer. But by understanding how computers think, we can tell computers what we would like them to do. If a computer is able to understand how humans think and give human-friendly communication accordingly, then the people that the computer communicates with do not have to understand how the computer thinks in order to understand the communication. Using Gordan & Hobbs' example [8], a computer which successfully adapts an anthropomorphic strategy should be able to make a human reasoning style statement — "Let me think through this a bit more slowly. I want to make sure I can remember it later." — instead of a computer reasoning style statement — "I am reducing the memory cache size by 128 megabytes to free up space on the hard drive". However, the computer does not have to use human reasoning all the way to achieve this response.

Anthropomorphic computing is fundamentally about giving computers an explicit awareness of the commonsense psychology model that is being applied to them, as well as the ability to reason about this model in support of human-computer interaction, as opposed to mirroring human-like reasoning in machines. Here, it is important to make sure that everyone participating in the interaction is using the same model. This model is an implicit commonsense theory of human psychology that applies to people. It needs to be explicitly represented in a formalism that is algorithm-friendly for computers, i.e. find a representation that is expressive enough for AI to reason about while being simple enough for human users to understand [8]. Horstman and Krämer [11] concluded on the basis of semi-structured interviews and a quantitative online study that not surprisingly people expect social robots to make their lives practically easier (e.g., by assisting with domestic or professional duties), but that they do not expect them helping with social activities. However, when asked about their preferences, the same participants bring up interpersonal and emotional skills more frequently than technological skills, and they express a yearning for empathetic social robots. In other words, and important for our point, at the moment there seems to be an interesting tension between on the one hand people's expectations about social robots (which focus on technological characteristics) and on the other hand people's preferences (which focus on emotional and social characteristics).

In order to successfully implement the anthropomorphic computing for co-operation and communication with people, we need to logically formalise folk psychology so that those folk inferential theory can be represented in the AI system. An informal attempt in finding axioms behind commonsense psychology is done by Smedslund in his book (1997) [24], "The structure of psychological common sense." Inspired by Smedslund's work, Gordan & Hobbs have been researching the logical formalisation of commonsense psychology in support of artificial intelligence that is similar to humans. In their recent book [8], they provide fourteen hundred first-order axioms of logic organised into twenty-nine theories and sixteen background theories, using formal logic to encode the entire breadth of psychological words and phrases. Future work might need to explore further to see if human treats intelligent AI robots as if their equals. For example, Weisma (2022) [30] has argued that human reasoners must consider two ontological questions in order to make sense of robots or any other entity:

– Which kind of thing is it?
– And what causal factors influence it?

Each question focuses on a different aspect of how robots are extraordinary —– though not exceptional —– entities for the human cognitive system. She provides a new theoretical framework for comprehending conceptual change at both the individual and cultural-historical levels by meditating on the dynamic interplay between these two questions. To summarise, research on formalising folk psychology theory has made decent progress but further research is still warranted to ensure such theory, built on human interaction dynamics, is applicable to the dynamics between humans and robots.

7 Verification of one's intuition with behavioural experiments

Relying on one's intuition to build new models and even use them as justification is nothing new. In fact, in the early development of psychology, introspection has been used as an inquiry method to study the mind. Wundt, for instance, developed a rigourous method in order to make it as reliable as possible. Notwithstanding these efforts, it remained criticized as subjective and unreliable. As a reaction, behaviourism not only promoted the use of experimental methods instead of introspection but also rejected the study of mind altogether as it is not directly observable and therefore not testable. Does it mean the endeavour of studying the mind is futile and we should only study behaviour? Not quite. These two camps pitch psychology as the study of behaviour against the study of the mind, but the quarrel has been "resolved" by the cognitive revolution in psychology which focuses on the study of cognitive mechanism, since it is assumed that behaviour is driven by an internal process. However, as we pointed out earlier, the study of cognitive models is different from the study of inferential theory, as the former can be seen as behaviour-driven and the latter can be seen as thought-driven. Following the cognitive revolution in psychology, it is possible to use human behavioural experiments to test models built on one's introspection and intuition [3].

While psychology has rejected the use of introspection (and intuition) as justification in what is right and what is wrong, other cognitive science fields such as linguistics and philosophy still rely upon it heavily as justification for their theory to this day. One point we would like to emphasize here is that there is nothing wrong with relying on one's own intuition to develop a model. In fact, our point of view is that it is actually beneficial to do so as it can give rise to ingenious and creative solutions to our scientific inquiry. However, it should be complemented with a method to verify those intuitive claims rather than saying that "it is obvious" or "it is commonsense" as justification or evidence.

In this regard, we can learn from the experience of how experimental psychology transformed pragmatics.

Over the course of its young history (about 40 years, see e.g., [7]), experimental pragmatics has experienced several changes. The study of pragmatic meaning by psychologists started around 1970. These psychologists came from different subfields, like those interested in developmental psychology or psycholinguistics, and used their experimental methods to investigate the use and understanding of language in context. In the spirit of cognitive science, collaborations between different disciplines turned out to be very fruitful. Linguists, logicians, and philosophers were for instance often the suppliers of inspiration and hypotheses, which then were experimentally tested by psychologists. This endeavour showed that the often heard claim that pragmatics is the wastebasket of linguistics and therefore an impossible topic for scientific investigation was ill-founded [7].

To conclude, since experimental pragmatics has successfully incentivized research from researcher's intuition to human behavioural experiment, it is reasonable to believe that we can still find human behavioural testable hypotheses on the current work of AI/XAI, even if the researchers who develop those models do not have those hypotheses in mind.

8 Conclusion: Human in the loop!

In this paper, we have seen the folk psychology concepts focus on the social aspects of human interactions. Humans have a tendency to anthropomorphise machines and AI, even though we know fully that they do not have intentions, because it helps us to navigate the world with the reasoning style that we are most familiar with. So it makes sense to make AI more "human-like". But how are we going to do that? The early work on human-computer interaction has already pointed out that having just a checklist — in our case, a checklist of what constitutes a good explanation in the eyes of humans — is not good enough, we also need to evaluate AI models with human subjects. It does not mean we — AI scientists with mainly computer science training — should not build models based on our own intuition. Nevertheless, we should be vigilant and avoid basing the justification for the AI model solely on one's intuition. Also, we should invite experimental psychologists to help us out with testing human judgements on our models if possible. After all, psychologists have been researching human behaviours for decades and they do have a successful track record in transforming their sister fields in cognitive science.

References

1. Carbonnelle, P., Deryck, M., Vennekens, J., et al.: An interactive consultant. In: BNAIC, Date: 2019/11/06-2019/11/08, Location: Bruxelles (2019)
2. Card, S., Moran, T., Newell, A.: The psychology of human-computer interaction (1983)

3. Davey, G., Sterling, C., Field, A., Sterling, C., Albery, I.: Complete psychology. Routledge (2014)
4. Denecker, M., Vennekens, J.: Building a knowledge base system for an integration of logic programming and classical logic. In: International Conference on Logic Programming. pp. 71–76. Springer (2008)
5. Durso, F.T., DeLucia, P., Jones, K.S.: Engineering psychology. The Corsini Encyclopedia of Psychology, John Wiley and Sons pp. 573–576 (2010)
6. Evans, J.S.B.: Dual process theory: Perspectives and problems. Dual process theory 2.0 pp. 137–155 (2017)
7. Gibbs Jr, R.W., Colston, H.L.: Pragmatics always matters: An expanded vision of experimental pragmatics. Frontiers in Psychology **11**, 1619 (2020)
8. Gordon, A.S., Hobbs, J.R.: A formal theory of commonsense psychology: How people think people think. Cambridge University Press (2017)
9. Grice, H.P.: Logic and conversation. In: Speech acts, pp. 41–58. Brill (1975)
10. Hansen, W.J.: User engineering principles for interactive systems. In: Proceedings of the November 16-18, 1971, fall joint computer conference. pp. 523–532 (1972)
11. Horstmann, A.C., Krämer, N.C.: Great expectations? relation of previous experiences with social robots in real life or in the media and expectancies based on qualitative and quantitative assessment. Frontiers in psychology **10**, 939 (2019)
12. Kirfel, L., Icard, T., Gerstenberg, T.: Inference from explanation. Journal of Experimental Psychology: General **151**(7), 1481 (2022)
13. Kirschner, F., Paas, F., Kirschner, P.A.: A cognitive load approach to collaborative learning: United brains for complex tasks. Educational psychology review **21**(1), 31–42 (2009)
14. Lombrozo, T.: The structure and function of explanations. Trends in cognitive sciences **10**(10), 464–470 (2006)
15. Malle, B.F.: How the mind explains behavior: Folk explanations, meaning, and social interaction. MIT press (2006)
16. Marr, D.: Vision: A computational investigation into the human representation and processing of visual information (1982)
17. Mercier, H., Sperber, D.: Why do humans reason? arguments for an argumentative theory. Behavioral and brain sciences **34**(2), 57–74 (2011)
18. Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. Artificial intelligence **267**, 1–38 (2019)
19. Miller, T., Howe, P., Sonenberg, L.: 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)
20. Moshman, D., Geil, M.: Collaborative reasoning: Evidence for collective rationality. Thinking & Reasoning **4**(3), 231–248 (1998)
21. Overton, J.A.: Explanation in science (2012)
22. Pearl, J., Mackenzie, D.: The book of why: the new science of cause and effect. Basic books (2018)
23. Schaffer, S.: Astronomers mark time: Discipline and the personal equation. Science in context **2**(1), 115–145 (1988)
24. Smedslund, J.: The structure of psychological common sense. Psychology Press (2013)
25. Sridharan, M., Gelfond, M., Zhang, S., Wyatt, J.: Reba: A refinement-based architecture for knowledge representation and reasoning in robotics. Journal of Artificial Intelligence Research **65**, 87–180 (2019)
26. Sridharan, M., Meadows, B.: Towards a theory of explanations for human–robot collaboration. KI-Künstliche Intelligenz **33**(4), 331–342 (2019)

27. Taylor, J.E.T., Taylor, G.W.: Artificial cognition: How experimental psychology can help generate explainable artificial intelligence. Psychonomic Bulletin & Review **28**(2), 454–475 (2021)
28. Turing, A.: Computing machinery and intelligence (1950)
29. Van Hertum, P., Dasseville, I., Janssens, G., Denecker, M.: The kb paradigm and its application to interactive configuration. Theory and Practice of Logic Programming **17**(1), 91–117 (2017)
30. Weisman, K.: Extraordinary entities: Insights into folk ontology from studies of lay people's beliefs about robots. In: Proceedings of the Annual Meeting of the Cognitive Science Society. vol. 44 (2022)
31. Wilson, K., Helton, W., Wiggins, M.: Cognitive engineering. Wiley Interdisciplinary reviews. Cognitive Science **4**(1), 17–31 (2012)