

Explanation in Artificial Intelligence: Insights from the Social Sciences

Tim Miller

*School of Computing and Information Systems
University of Melbourne, Melbourne, Australia
tmiller@unimelb.edu.au*

Abstract

There has been a recent resurgence in the area of explainable artificial intelligence as researchers and practitioners seek to provide more transparency to their algorithms. Much of this research is focused on explicitly explaining decisions or actions to a human observer, and it should not be controversial to say that, if these techniques are to succeed, the explanations they generate should have a structure that humans accept. However, it is fair to say that most work in explainable artificial intelligence uses only the researchers' intuition of what constitutes a 'good' explanation. There exists vast and valuable bodies of research in philosophy, psychology, and cognitive science of how people define, generate, select, evaluate, and present explanations. This paper argues that the field of explainable artificial intelligence should build on this existing research, and reviews relevant papers from philosophy, cognitive psychology/science, and social psychology, which study these topics. It draws out some important findings, and discusses ways that these can be infused with work on explainable artificial intelligence.

Keywords: Explanation, Explainability, Interpretability, Explainable AI, Transparency

Contents

1	Introduction	3
1.1	Scope	4
1.2	Major Findings	5
1.3	Outline	6
1.4	Example	6
2	Philosophical Foundations — What Is Explanation?	7
2.1	Definitions	7
2.1.1	Causality	7
2.1.2	Explanation	10
2.1.3	Explanation as a Product	11
2.1.4	Explanation as a Process	11
2.1.5	Interpretability	12
2.2	Why People Ask for Explanations	12

2.3	Contrastive Explanation	13
2.4	Types and Levels of Explanation	14
2.5	Structure of Explanation	15
2.6	Explanation and Artificial Intelligence	17
2.6.1	Causal Attribution \neq Causal Explanation	17
2.6.2	Contrastive Explanation	18
2.6.3	Explanatory Tasks and Levels of Explanation	18
2.6.4	Model of Self	19
2.6.5	Structure of Explanation	20
3	Social Attribution — How Do People Explain Behaviour?	21
3.1	Definitions	21
3.2	Intentionality and Explanation	22
3.3	Beliefs, Desires, Intentions, and Traits	23
3.3.1	Malle’s Conceptual Model for Social Attribution	23
3.4	Individual vs. Group Behaviour	25
3.5	Norms and Morals	26
3.6	Social Attribution and Artificial Intelligence	27
3.6.1	Folk Psychology	27
3.6.2	Malle’s Models	28
3.6.3	Collective Intelligence	28
3.6.4	Norms and Morals	28
4	Cognitive Processes — How Do People Select and Evaluate Explanations?	29
4.1	Causal Connection, Explanation Selection, and Evaluation	29
4.2	Causal Connection: Counterfactuals and Mutability	31
4.2.1	Abnormality	31
4.2.2	Temporality	32
4.2.3	Controllability and Intent	32
4.2.4	Social Norms	33
4.2.5	Explanatory Modes	33
4.2.6	Principles Connections in Formal Explanations	34
4.3	Explanation Selection	35
4.3.1	Facts and Foils	35
4.3.2	Abnormality	36
4.3.3	Intentionality and Functionality	36
4.3.4	Necessity	37
4.3.5	Responsibility	37
4.3.6	Preconditions, Failure, and Intentions	38
4.4	Explanation Evaluation	39
4.4.1	Coherence, Simplicity, and Generality	39
4.4.2	Truth and Probability	39
4.4.3	Goals and Explanatory Mode	41
4.5	Cognitive Processes and Artificial Intelligence	41
4.5.1	Mutability and Computation	41
4.5.2	Abnormality	41

4.5.3	Intentionality and Functionality	42
4.5.4	Perspectives and Controllability	42
4.5.5	Evaluation of Explanations	42
5	Social Explanation — How Do People Communicate Explanations?	43
5.1	Explanation as Conversation	43
5.1.1	Logic and Conversation	44
5.1.2	Relation & Relevance in Explanation Selection	45
5.1.3	Linguistic structure	47
5.1.4	Argumentation and Explanation	47
5.2	Dialogue	48
5.3	Social Explanation and Artificial Intelligence	49
5.3.1	Conversational Model	49
5.3.2	Model of Other	50
5.3.3	Implicature	50
5.3.4	Dialogue	50
5.3.5	Dilution	51
6	Conclusions	51

1. Introduction

Recently, the notion of *explainable artificial intelligence* has seen a resurgence, after having slowed since the burst of work on explanation in expert systems over three decades ago; for example, see Chandrasekaran et al. [12], [125], and Buchanan and Shortliffe [8]. Sometimes abbreviated XAI (eXplainable artificial intelligence), the idea can be found in grant solicitations [18] and in the popular press [102]. This resurgence is driven by evidence that many AI applications have limited take up, or are not appropriated at all, due to a *lack of trust* on behalf of their users [123, 70]. The running hypothesis behind the explainable artificial intelligence is that by designing and implementing intelligence agents that are *transparent*, users will be better equipped to understand and therefore trust the intelligent agents [97, 13, 40].

While there are many ways to increase trust and transparency of intelligent agents, two complementary approaches will form part of many trusted autonomous systems: (1) generating decisions¹ in which one of the criteria taken into account during the computation is how well a human could understand the decisions in the given context, which is often called *interpretability* or *explainability*; and (2) explicitly explaining decisions to people, which we will call *explanation*. Applications of explanation are considered in many sub-fields of artificial intelligence, such as justifying autonomous agent behaviour [97, 40], debugging of machine learning models [62], explaining medical decision-making [26], and explaining predictions of classifiers [117].

If we want to design, and implement intelligent agents that are truly capable of providing explanations to people, then it is fair to say that models of how humans explain

¹We will use *decision* as the general term to encompass outputs from AI systems, such as categorisations, action selection, etc.

decisions and behaviour to each other are a good way to start analysing the problem. Despite the recent resurgence of explanation and interpretability in AI, it is fair to say that most of the research and practice in this area uses the researchers’ intuitions of what constitutes a ‘good’ explanation [100]. However, as argued by Miller et al. [100], this could lead to failure. The very experts who understand decision-making models the best are not in the right position to judge the usefulness of explanations to lay users — a phenomenon that Miller et al. refer to (paraphrasing Cooper [17]) as “the inmates running the asylum”. Therefore, a strong understanding of how people define, generate, select, evaluate, and present explanations seems almost essential.

In the fields of philosophy, cognitive psychology/science, and social psychology, there is a vast and mature body of work that studies these exact topics. For millennia, philosophers have asked the questions about what constitutes an explanation, what is the function of explanations, and what are their structure. For over 50 years, cognitive and social psychologists have analysed how people attribute and evaluate the social behaviour of others in physical environments. For over two decades, cognitive psychologists and scientists have investigated how people generate explanations and how they evaluate their quality.

I argue here that there is considerable scope to infuse this valuable body of research into explainable AI. Building intelligent agents capable of explanation is a challenging task, and approaching this challenge in a vacuum considering only the computational problems will not solve the greater problems of trust and transparency in AI. Further, while some recent work builds on the early findings on explanation in expert systems, that early research was undertaken prior to much of the work on explanation in social science. I contend that newer theories should now form the basis of explainable AI — although there is still a lot to learn from early work in explainable AI around design and implementation.

Finding relevant articles to apply to a particular AI problem is a daunting and time-consuming task. There are several hundred articles on the topic of explanation. It is not reasonable to expect every artificial intelligence researcher or practitioner to navigate this vast array of work. This is made even more difficult by the fact that the style of research and writing in social sciences differs to that of computer science and engineering, and therefore, these articles can be less accessible to computer scientists and engineers than to their intended audience.

This paper aims to promote the inclusion of this existing research into the field of explanation in AI. As part of this work, over 200 papers on explanation were surveyed from social science venues. A smaller subset of these were chosen to be presented in this paper, based on their relevance to the current understanding of the topic in social science, and on their relevance to explanation in AI. The paper should be used to refer the reader to the relevant work in social sciences for their particular domain of interest.

The paper presents relevant current theories on explanation, describes, in many cases, the relevant experimental evidence supporting these theories, and presents ideas on how this work can be infused into explainable AI.

1.1. Scope

The scope of this article threefold:

- *Survey*: To survey and review relevant articles on the philosophical, cognitive, and social foundations of explanation, with an emphasis on ‘everyday’ explanation.
- *Everyday explanation*: To focus on ‘everyday’ explanations as a tool and process for an agent, who we call the *explainer*, to explain decisions made by *itself or another agent* to a *person*, who we call the *explainee*. ‘Everyday’ explanations are the explanations of why particular facts (events, properties, decisions, etc.) occurred, rather than explanations of more general relationships, such as those seen in scientific explanation. We justify this focus based on the observation from AI literature that trust is lost when users cannot understand traces of observed behaviour or decisions [123, 97], rather than trying to understand and construct generalised theories. Despite this, everyday explanations also sometimes refer to generalised theories, as we will see later in Section 2, so scientific explanation is relevant, and some work from this area is surveyed in the paper.
- *Relationship to AI*: To draw important points from relevant articles to some of the different sub-fields of artificial intelligence.

The following topics are considered *out of scope* of this article:

- *Causality*: While causality is important in explanation, this paper is not a survey on the vast work on causality. I review the major positions in this field insofar as they relate to the relationship with models of explanation.
- *Explainable AI*: This paper is not a survey on existing approaches to explanation or interpretability in AI, except those that directly contribute to the topics in scope or to some of our discussions.

1.2. Major Findings

As part of this review, I highlight three major findings from the surveyed literature that I believe are important for explainable AI, but which I believe most research and practitioners in artificial intelligence are currently unaware:

1. Explanations are *contrastive* — they are sought in response to particular *counterfactual cases*, which are termed *foils* in this paper. That is, people do not ask why event *P* happened, but rather why event *P* happened *instead of* some event *Q*. This has important social and computational consequences for explainable AI. Sections 2 and 4, models of how people provide contrastive explanations are reviewed.
2. Explanations are *selected* — people rarely, if ever, expect an explanation that consists of an actual and complete cause of an event. Humans are adept at selecting one or two causes from a sometimes infinite number of causes to be *the* explanation. In Section 4, models of how people select explanations, including how this relates to contrast cases, are reviewed.
3. Explanations are *social* — they are a transfer of knowledge, presented as part of a conversation² or interaction, and are thus presented relative to the explainer’s

²Note that this does not imply that explanations must be given in natural language, but implies that explanation is a social interaction between the explainer and the explainee.

beliefs about the explainee’s beliefs. In Section 5, models of how people interact regarding explanations are reviewed.

These three points all converge around a single point: explanations are not just the presentation of causes (*causal attribution*). While an event may have many causes, often the explainee cares only about a small subset (relevant to the contrast case), the explainer selects a subset of this subset (based on several different criteria), and explainer and explainee may interact and argue about this explanation.

I assert that, if we are to build truly explainable AI, especially intelligent systems that are able to offer explanations, then these three points are imperative in many applications.

1.3. Outline

The outline of this paper is as follows. Section 1.4 presents a motivating example of an explanatory agent that is used throughout the paper. Section 2 presents the philosophical foundations of explanation, defining what explanations are, what they are not, how to relate to causes, their meaning and their structure. Section 3 focuses on one specific type of explanation — those relating to human or social behaviour, while Section 4 surveys work on how people generate and evaluate explanations more generally; that is, not just social behaviour. Section 5 describes research on the dynamics of interaction in explanation between explainer and explainee. Section 6 concludes and highlights several major challenges to explanation in AI.

1.4. Example

This section presents a simple example, which is used to illustrate many important concepts through this paper. It is of a hypothetical system that categorises images of arthropods into several different types, based on certain physical features of the arthropods, such as number of legs, number of eyes, number of wings, etc. The algorithm is assumed to have been trained on a large set of valid data and is highly accurate. It is used by entomologists to do automatic classification of their research data. Table 1 outlines a simple model of the features of arthropods for illustrative purposes. An explanation function is available for the arthropod system.

Type	No. Legs	Stinger	No. Eyes	Compound Eyes	Wings
Spider	8	✗	8	✗	0
Beetle	6	✗	2	✓	2
Bee	6	✓	5	✓	4
Fly	6	✗	5	✓	2

Table 1: A simple lay model for distinguishing common arthropods.

Now, consider the idealised and simple dialogue between a human user and ‘ExplA-gent’, who is the interactive explanation agent, outlined in Table 1. This dialogue is not intended to be realistic, but is merely illustrative of how a particular explanatory agent may interact: responding to posed questions, using mixed modalities — in this case, language and visual images — and being able to answer a range of questions about

its decision making. This example shows different types of questions being posed, and demonstrates that the explanatory agent will need to keep track of the state of the explanation; for example, by noting what it has already told the explainee, and may have to infer what the explainee has inferred themselves.

Person:	<i>“Why is image J labelled as a Spider instead of a Beetle?”</i>
ExplAgent:	<i>“Because the arthropod in image J has eight legs, consistent with those in the category Spider, while those in Beetle have six legs.”</i>
Person:	<i>“Why did you infer that the arthropod in image J had eight legs instead of six?”</i>
ExplAgent:	<i>“I counted the eight legs that I found, as I have just highlighted on the image now.”</i> (ExplAgent shows the image with the eight legs counted).
Person:	<i>“How do you know that spiders have eight legs?”</i>
ExplAgent:	<i>“Because in the training set I was trained on, almost all animals with eight legs were labelled as Spider.”</i>
Person:	<i>“But an octopus can have eight legs too. Why did you not classify image J as an octopus?”</i>
ExplAgent:	<i>“Because my function is only to classify arthropods.”</i>

Figure 1: Example Explanation Dialogue between a Person and an Explanation Agent

We will refer back to this example throughout the paper and link different parts of work the different parts of the dialogue above.

2. Philosophical Foundations — What Is Explanation?

“To explain an event is to provide some information about its causal history. In an act of explaining, someone who is in possession of some information about the causal history of some event — explanatory information, I shall call it — tries to convey it to someone else.” – Lewis [69, p. 217].

In this section, we outline some foundational work in explanation, which helps to define causal explanation and how it differs from other concepts, such as causal attribution and interpretability/explainability.

2.1. Definitions

There are several related concepts in explanation, which seem to be used interchangeably between authors and also within articles, often demonstrating some conflation of the terms. In particular, this section describes the difference between causal attribution and causal explanation. We will also briefly touch on the difference between explanation and interpretability.

2.1.1. Causality

The idea of causality has attracted much work, and there are several different accounts of what constitutes a *cause* of an event or property. The various definitions of causation can be broken into two major categories: *dependence theories* and *transference theories*.

Causality and Counterfactuals. Hume and Beauchamp [54, Section VII] is credited with deriving what is known as the *regularity theory* of causation. This theory states that there is a cause between two types of events if events of the first type are always followed by events of the second. However, as argued by Lewis [68], the definition due to Hume and Beauchamp is in fact about *counterfactuals*, rather than dependence alone. Hume and Beauchamp argues that the co-occurrence of events C and E , observed from experience, do not give causal information that is useful. Instead, the cause should be understood relative to an imagined, counterfactual case: event C is said to have caused event E if, under some hypothetical counterfactual case the event C did not occur, E would not have occurred. This definition has been argued and refined, and many definitions of causality are based around this idea in one way or another; c.f. Lewis [68], Hilton [47].

This *classical counterfactual* model of causality is well understood but competing definitions exist. *Interventionist* theories of causality [142, 35] state that event C can be deemed a cause of event E if and only if any change to event E can be brought about solely by intervening on event C . *Probabilistic* theories, which are extensions of interventionist theories, state that event C is a cause of event E if and only if the occurrence of C increases the probability of E occurring [96].

Transference theories [3, 25, 21], on the other hand, are not defined on dependence, but instead describe physical causation as the transference of energy between objects. In short, C causes E if even E represents an unnatural change of energy in the object of E , that object is in contact with the object that causes E , and there is some *quantity* of energy transferred.

Our aim here is not to survey and discuss definitions of causality, however, it is pertinent to note that the dependence theories all focus around the concept of *counterfactuals*: the state of affairs that *would have resulted* from some event that *did not occur*. Even transference theories, which are not explicitly defined as counterfactual, consider that causation is an *unnatural* transference of energy to the receiving object, implying what would have been otherwise. As such, the notion of counterfactuals is important in causality.

Necessary and Sufficient Causes. Kelley [60] proposes a taxonomy of causality in social attribution, but which has more general applicability, and noted that there are two main types of *causal schemata* for causing events: *multiple necessary causes* and *multiple sufficient causes*. The former defines a schema in which a set of events are all necessary to cause the event in question, while the latter defines a schema in which there are multiple possible ways to cause the event, and only one of these is required. Clearly, these can be interleaved; e.g. causes C_1 , C_2 , and C_3 for event E , in which C_1 is necessary and either of C_2 or C_3 are necessary, while both C_2 and C_3 are sufficient to cause the compound event (C_2 or C_3).

Internal and External Causes. Heider [41], the grandfather of causal attribution in social psychology, argues that causes fall into two camps: internal and external. Internal causes of events are those due to the characteristics of an actor, while external causes are those due to the specific situation or the environment. Clearly, events can have causes that mix both. However, the focus of work from Heider was not on causality in general, but on *social attribution*, or the *perceived* causes of behaviour. That is, how people attribute

the behaviour of others. Nonetheless, work in this field, as we will see in Section 3, builds heavily on counterfactual causality.

Causal Chains. In causality and explanation, the concept of *causal chains* is important. A causal chain is a path of causes between a set of events, in which a cause from event C to event E indicates that C must occur before E . Any events without a cause are *root causes*.

Hilton et al. [52] define five different types of causal chain, outlined in Table 2, and note that different causal chains are associated with different types of explanations.

Type	Description	Example
Temporal	Distal events do not constraint proximal events. Events can be switched in time without changing the outcome	A and B together cause C ; order of A and B is irrelevant; e.g. two people each flipping a coin win if both coins are heads; it is irrelevant who flips first.
Coincidental	Distal events do not constraint proximal events. The causal relationships holds in a particular case, but not in general.	A causes B this time, but the general relationship does not hold; e.g. a person smoking a cigarette causes a house fire, but this does not generally happen.
Unfolding	Distal events strongly constrain proximal events. The causal relationships hold in general and in this particular case and cannot be switched.	A causes B and B causes C ; e.g. switching a light switch causes an electric current to run to the light, which causes the light to turn on
Opportunity chains	The distal event <i>enables</i> the proximal event.	A enables B , B causes C ; e.g. installing a light switch enables it to be switched, which causes the light to turn on.
Pre-emptive	Distal precedes proximal and prevents the proximal from causing an event.	B causes C , A would have caused C if B did not occur; e.g. my action of unlocking the car with my remote lock would have unlocked the door if my wife had not already unlocked it with the key.

Table 2: Types of Causal Chains according to Hilton et al. [52].

People do not need to understand a complete causal chain to provide a sound explanation. This is evidently true: explanations about physical events can refer back to events that occurred during the Big Bang, but nonetheless, most adults can explain to a child why a bouncing ball eventually stops.

Formal Models of Causation. While several formal models of causation have been proposed, such as those based on conditional logic [30, 68], the model of causation that I believe would be of interest to many in artificial intelligence is the formalisation of causality by Halpern and Pearl [35]. This is a general model that should be accessible to anyone with a computer science background, has been adopted by philosophers and

psychologists, and is accompanied by many additional results, such as an axiomatisation [34] and a series of articles on complexity analysis [22], including an analysis of those cases that are tractable [23].

Halpern and Pearl [35] define a model-based approach using *structural causal models* over two sets of variables: *exogenous* variables, whose values are determined by factors external to the model, and *endogenous* variables, whose values are determined by relationships with other (exogenous or endogenous) variables. Each endogenous variable has a function that defines its value from other variables. A *context* is an assignment from each variable to a unique value for that variable. Intuitively, a context represents a ‘possible world’ of the model. A model/context pair is called a *situation*. Given this structure, Halpern and Pearl define a *actual cause* of an event $X = x$ (that is, endogenous variable X receiving the value x) as a set of events E (each of the form $Y = y$) such that (informally) the following three criteria hold:

AC1 Both the event $X = x$ and the cause E are true in the actual situation.

AC2 If there was some *counterfactual* values for the variables of the events in E , then the event $X = x$ would not have occurred.

AC3 E is minimal — that is, there are no irrelevant events in the case.

A *sufficient cause* is simply a non-minimal actual cause; that is, it satisfies the first two items above.

We will return later to this model in Section 5.1.2 to discuss Halpern and Pearl’s model of explanation.

2.1.2. Explanation

The word “explanation” is both a verb and a noun³, and is thus both a process and a product, as noted by Lombrozo [73]. However, I argue that there are actually *two* processes in explanation, as well as the product:

1. *Cognitive process* — The process of abductive inference for determining an explanation for a given event, called the *explanandum*, in which the causes for the event are identified, perhaps in relation to a particular counterfactual cases, and a subset of these causes is selected as *the explanation* (or *explanans*).
2. *Product* — The explanation that results from the cognitive process is the *product* of the cognitive explanation process.
3. *Social process* — The process of transferring knowledge between explainer and explainee, generally an interaction between a group of people, in which the goal is that the explainee has enough information to understand the *causes* of the event; although other types of goal exists, as we discuss later.

³Of course, this also means it can be used as an adjective as well, but to simplify the argument, we treat it as a verb and noun only.

But what constitutes an explanation? This question has created a lot of debate in philosophy, but accounts of explanation both philosophical and psychology stress the importance of causality in explanation — that is, an explanation refers to a cause [118, 142, 76, 36]. There are, however, definitions of non-causal explanation [29], such as explaining ‘what happened’ or explaining what was meant by a particular remark [140]. These definitions out of scope in this paper, and they present a different set of challenges to explainable AI.

2.1.3. *Explanation as a Product*

We take the definition that an ‘everyday’ explanation is an answer to a *why-question* [19, 104, 69, 71].

According to Bromberger [7], a why-question is a combination of a *whether-question*, preceded by the word ‘why’. A whether-question is an interrogative question whose correct answer is either ‘yes’ or ‘no’. The *presupposition* within a why-question is the fact referred to in the question that is under explanation, expressed as if it were true (or false if the question is a negative sentence). For example, the question “*why did they do that?*” is a why-question, with the inner whether-question being “*did they do that?*”, and the presupposition being “*they did that*”. However, as we will see in Section 2.3, why-questions are structurally more complicated than this: they are *contrastive*.

As noted by Overton [104], explanations can be answers to other types of questions, particularly some *how-questions*. However, Overton argues that for the purposes of explanation, how-questions can be turned into why-questions; e.g. “*how does copper turn green when exposed to air?*” can be re-phrased to “*why does copper turn green when exposed to air?*”. Other types of questions such as “*How tall is that flagpole?*” or “*What happened?*” cannot, but these are asking for a factual account, rather than a counterfactual causal explanation. Clearly, we can also consider ‘*for what*’-questions as why-questions as well. While the definitions from Bromberger [7] and Overton [104] do not seem to cover all why-questions, they serve as a sufficient definition for ‘everyday’ explanations.

Section 2.4 discusses answers to why-questions in more detail.

2.1.4. *Explanation as a Process*

From the dynamic perspective, explanation is the processes of deriving an explanation product; that is, it ‘fills in gaps’ of knowledge in order to attribute causes to an event [15]. As noted already, this process can be either *cognitive*, in which a person generates explanation products to identify cause for themselves, or *social*, in which an explainer transfers knowledge using explanation products for the explainee to help them identify causes.

In Section 4, we discuss the cognitive science view of explanation and in particular, more precise definitions of why-questions. In particular, throughout Sections 4–5, explanations themselves are not just the causes of an event. Instead, explanations are typically just one or two (of the possibly infinite number) causes of an event, and how and why these are selected is not arbitrary. It is this property that is the main contrast between causal attribution and causal explanation.

2.1.5. Interpretability

Here, we briefly address the distinction between *interpretability*, *explainability* and *explanation*, as used in this article; and as it seems to be used in artificial intelligence. Section 2.1.2 defines ‘explanation’, but there are the broader terms that people refer to as ‘interpretability’ or ‘explainability’. The terms ‘interpretability’ and ‘explainability’ invoke the idea of a measure, in that one agent can be seen to have a higher interpretability than another.

Lipton [72] provides a taxonomy of the desiderata and methods for interpretable AI. Although his arguments discuss only machine learning, the ideas can be generalised to other AI models. This paper adopts [72]’s assertion that explanation is post-hoc interpretability. I use a simple definition of *interpretability* of a model as: the degree to which an observer can understand the cause of a decision. Post-hoc explanation is thus one mode in which an observer may obtain understanding, but clearly, there are additional modes that one can adopt, such as making decisions that are inherently easier to understand. I equate ‘interpretability’ with ‘explainability’.

It is important to understand the similarities and differences between these terms as one reads this article, because some related research discussed is relevant to explanation only, in particular, Section 5, which discusses how people present explanations to one another; while other sections, in particular Sections 3 and 4 discuss how people generate and evaluate explanations, and explain behaviour of others, so are broader and can be used to create more interpretable agents.

2.2. Why People Ask for Explanations

There are many reasons that people may ask for explanations. Curiosity is one primary criterion that humans use, but other pragmatic reasons include examination — for example, a teacher asking her students for an explanation on an exam for the purposes of testing the students’ knowledge on a particular topic; and scientific explanation — asking why we observe a particular environmental phenomenon.

In this paper, we are interested in explanation in AI, and thus our focus is on how intelligent agents can explain their decisions. As such, this section is primarily concerned with why people ask for ‘everyday’ explanations of why *specific* events occur, rather than explanations for general scientific phenomena or.

It is clear that primary function of explanation is to facilitate *learning* [73, 141]. Via learning, we obtain better models of how particular events or properties come about, and we are able to use these models to our advantage. Heider [41] states that people look for explanations to improve their understanding of someone or something so that they can derive stable model that can be used for prediction and control. This hypothesis is backed up by research suggesting that people tend to ask questions about events or observations that they consider abnormal or unexpected from their own point of view [53, 49, 44]. Thus, when someone observes an event or property that violates their understanding, they ask why-questions to correct and improve that model.

Lombrozo [73] argues that explanations have a role in inference learning precisely *because* they are explanations, not necessarily just due to the causal information they reveal. First, explanations provide somewhat of a ‘filter’ on the causal beliefs of an event. Second, prior knowledge is changed by giving explanations; that is, by asking someone to provide an explanation as to whether a particular property is true or false,

the explainer changes their perceived likelihood of the claim. Third, explanations that offer fewer causes and explanations that explain multiple observations are considered more believable and more valuable; but this does not hold for causal statements.

Malle [80, Chapter 3], who gives perhaps the most complete discussion of everyday explanations in the context of explaining social action/interaction, argues that people ask for explanations for two reasons:

1. *To find meaning*: to reconcile the contradictions or inconsistencies between elements of our knowledge structures.
2. *To manage social interaction*: to create a *shared meaning* of something, and to change others' beliefs & impressions, their emotions, or to influence their actions.

Creating a shared meaning is important for explanation in AI. In many cases, an explanation provided by an intelligent agent will be precisely to do this — to create a shared understanding of the decision that was made between itself and a human observer, at least to some partial level.

Lombrozo [73] notes that explanations have several functions other than the transfer of knowledge, such as persuasion or assignment of blame; and that in some cases of social explanation, the goals of the explainer and explainee may be different. With respect to explanation in AI, persuasion is surely of interest: if the goal of an explanation from an intelligent agent is to generate trust from a human observer, then persuasion that a decision is the correct one could in some case be considered more important than actually transferring the true cause. In this case, the goals of the explainer (to generate trust) is different to that of the explainee (to understand a decision).

2.3. Contrastive Explanation

“The key insight is to recognise that one does not explain events per se, but that one explains why the puzzling event occurred in the target cases but not in some counterfactual contrast case.” — Hilton [48, p. 67].

I will dedicate a subsection to discuss one of the most important findings in the philosophical and cognitive science literature from the perspective of explainable AI: *contrastive explanation*. Research shows that people do not explain the causes for an event *per se*, but explain the cause of an event *relative to some other event* that did not occur; that is, an explanation is always of the form “*Why P rather than Q?*”, in which *P* is the target event and *Q* is a counterfactual contrast case that did not occur, even if the *Q* is implicit in the question. This is called *contrastive explanation*.

Some authors refer to *Q* as the *counterfactual case* [77, 44, 53]. However, it is important to note that this is not the same counterfactual that one refers to when determining causality (see Section 2.1.1). For causality, the counterfactuals are hypothetical ‘non-causes’ in which the event-to-be-explained does not occur — that is a counterfactual to cause *C* —, whereas in contrastive explanation, the counterfactuals are hypothetical outcomes — that is, a counterfactual to event *E*.

Lipton [71] refers to the two cases, *P* and *Q*, as the *fact* and the *foil* respectively; the *fact* being the event that did occur, and the *foil* being the event that did not. To avoid confusion, throughout the remainder of this paper, we will adopt this terminology and

use *counterfactual* to refer to the hypothetical case in which the cause C did not occur, and *foil* to refer to the hypothesised case Q that was expected rather than P .

Authors in this area argue that all why-questions ask for contrastive explanations, even if the foils are not made explicit [71, 53, 44, 48, 78, 77], and that humans are well trained at detecting the foil from the language and tone. For example, given the question, “*Why did Elizabeth open the door?*”, there are many, possibly an infinite number, of foils; e.g. “*Why did Elizabeth open the door, rather than leave it closed?*”, “*Why did Elizabeth open the door rather than the window?*”, or “*Why did Elizabeth open the door rather than Michael opening it?*”. These different contrasts have different explanations, and there is no inherent one that is certain to be the foil for this question. The negated presupposition *not(Elizabeth opens the door)* refers to an entire class of foils, including all those listed already.

It is important that the explainees understand the counterfactual case [44]. For example, given the question “*Why did Elizabeth open the door?*”, the answer “*Because she was hot*” is a good answer if the foil is Elizabeth leaving the door closed, but not a good answer if the foil is “*rather than turning on the air conditioning*”, because the fact that Elizabeth is hot explains both the fact and the foil.

The idea of contrastive explanation should not be controversial if we accept the argument outlined in Section 2.2 that people ask for explanations about events or observations that they consider abnormal or unexpected from their own point of view [53, 49, 44]. In such cases, people expect to observe a particular event, but then observe another, with the observed event being the fact and the expected event being the foil.

Lipton [71] argues that contrastive explanations between a fact P and a foil Q are, in general, easier to derive than ‘complete’ explanations for P . For example, consider the arthropod classification algorithm in Section 1.4. To be a beetle, an arthropod must have six legs, but this does not cause an arthropod to be a beetle – other causes are necessary. Lipton contends that we could answer a question such as “*Why is image J labelled as a Beetle instead of a Spider?*” by citing the fact that the arthropod in the image has six legs. We do not need information about eyes, wings, or stingers to answer this, whereas to explain why image J is a spider in a non-contrastive way, we must cite all causes.

The hypothesis that all causal explanations are counterfactual is not merely philosophical. In Section 4, we see several bodies of work supporting this, and these provide more detail as to how people select and evaluate explanations based on the contrast between fact and foil.

2.4. Types and Levels of Explanation

The type of explanation provided to a question is dependent on the particular question asked; for example, asking why some event occurred is different to asking under what circumstances it *could have* occurred; that is, the actual vs. the hypothetical [118]. However, for the purposes of answering why-questions, we will focus on a particular subset of philosophical work in this area.

Aristotle’s *Four Causes* model, also known as the *Modes of Explanation* model, continues to be foundational for cause and explanation. Aristotle proposed an analytic scheme, classed into four different elements, that can be used to provide answers to why-questions [37]:

1. *Material*: The substance or material of which something is made. For example, rubber is a material cause for a car tyre.
2. *Formal*: The form or properties of something that make it what it is. For example, being round is a formal cause of a car tyre. These are sometimes referred to as *categorical* explanations.
3. *Efficient*: The proximal mechanisms of the cause something to change. For example, a tyre manufacturer is an efficient cause for a car tyre. These are sometimes referred to as *mechanistic* explanations.
4. *Final*: The end or goal of something. Moving a vehicle is an efficient cause of a car tyre. These are sometimes referred to as *functional* or *teleological* explanations.

A single why-question can have explanations from any of these categories. For example, consider the question: “*Why does this pen contain ink?*”. A material explanation is based on the idea that the pen is made of a substance that prevents the ink from leaking out. A formal explanation is that it is a pen and pens contain ink. A efficient explanation is that there was a person who filled it with ink. A final explanation is that pens are for writing, and so require ink.

Several other authors have proposed models similar to Aristotle’s, such as Dennett [19], who proposed that people take three *stances* towards objects: *physical*, *design*, and *intention*; and Marr [87], building on earlier work with Poggio [88], who define the *computational*, *representational*, and *hardware* levels of understanding for computational problems.

Kass and Leake [58] define a categorisation of explanations of anomalies into three types: (1) *intentional*; (2) *material*; and (3) *social*. The intentional and material categories correspond roughly for Aristotle’s final and material categories, however, the *social* category does not correspond to any particular category in the models of Aristotle, Marr [87], or Dennett [19]. The social category refers to explanations about human behaviour that is not intentionally driven. Kass and Leake give the example of an increase in crime rate in a city, which, while due to intentional behaviour of individuals in that city, is not a phenomenon that can be said to be intentional. While individual crimes are committed with intent, it cannot be said that the individuals had the intent of increasing the crime rate — that is merely an effect of the behaviour of a group of individuals.

2.5. Structure of Explanation

As we saw in Section 2.1.2, causation is a major part of explanation. Earlier accounts of explanation from Hempel and Oppenheim [43] argued for logically deductive models of explanation. Kelley [59] subsequently argued instead that people consider *co-variation* in constructing explanations, and proposed a *statistical* model of explanation based ANOVA. However, subsequent experimental research uncovered many problems with these models, and currently, both the deductive and statistical models of explanation are no longer considered valid theories in most camps [82].

Overton [106, 105] defines a model of scientific explanation, which, despite having different goals to ‘everyday’ explanation, provides a useful model of explanation in AI. In particular, Overton [105] defines the *structure* of explanations. Overton defines five

categories of properties or objects that are explained in science: (1) *theories*: sets of principles that form building blocks for models; (2) *models*: an abstraction of a theory that represents the relationships between kinds and their attributes; (3) *kinds*: an abstract universal class that supports counterfactual reasoning; (4) *entities*: an instantiation of a kind; and (5) *data*: statements about activities (e.g. measurements, observations). The relationships between these is shown in Figure 2.

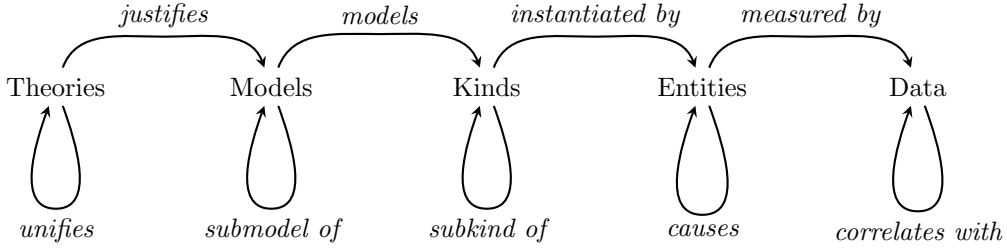


Figure 2: Overton’s five categories and four relations in scientific explanation, reproduced from Overton [105, p. 54, Figure 3.1]

From these categories, Overton [105] provides a crisp definition of the structure of scientific explanations. As part of this structure, Overton notes that explanations answer questions in part by relating the properties of different categories. He argues that an explanation of a phenomenon at one level by appealing to another level must refer to all intermediate levels. For example, a categorical explanation can be given by noting that it is measured from a particular entity, which is of a particular kind, and all of those kinds have a particular property that explains the observation. Figure 3 shows the structure of a *theory-data explanation*, which Overton argues is the most complex because on this model, *Theory* and *Data* have the longest chain of relationships between them.

This model provides a solid foundation for explanation in AI. For example, consider our arthropod classification algorithm from Section 1.4. The explanation agent could explain that a particular piece of *Data* — the number of legs being eight — is observed in this particular *Entity* — the specific instance of in this image —, and *Entities* of the *Kind* spider have eight legs according to a *Model*, while *Entities* of the *Kind* beetles have only six legs. In this example, the explanation is constructed by appealing to the *Model* of insects. This could, in turn, appeal to a particular *Theory* that underlies that *Model*.

With respect to social explanation, Malle [80] argues that social explanation is best understood as consisting of three layers:

1. Layer 1: A conceptual framework that outlines the assumptions people make about human behaviour and explanation.
2. Layer 2: The psychological processes that are used to construct explanations.
3. Layer 3: Language layer that specifies the type of linguistic structures people use in giving explanations.

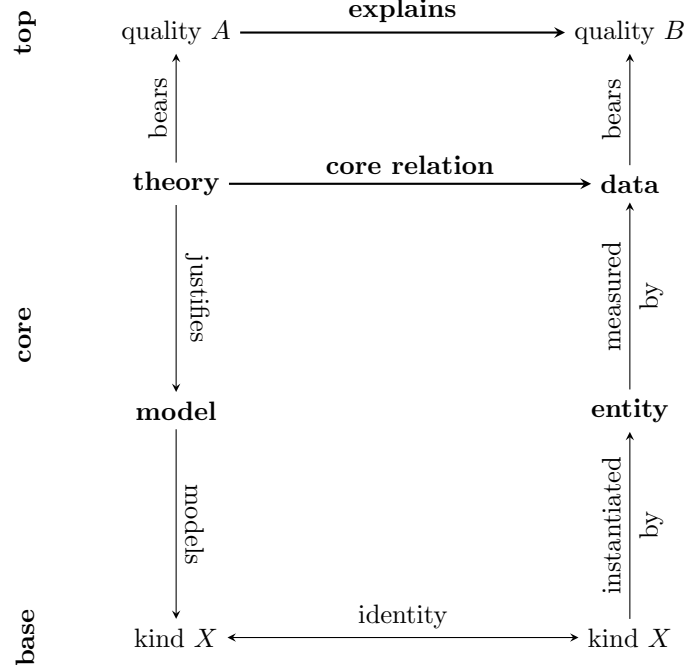


Figure 3: Overton’s general structure of a theory-data explanation, reproduced from Overton [105, p. 54, Figure 3.2])

I will present Malle’s views of these three layers in more detail in the section on social attribution (Section 3), cognitive processes (Section 4, and social explanation (Section 5). This work is collated into Malle’s 2004 book [80].

2.6. Explanation and Artificial Intelligence

This section presents some ideas on how the philosophical work outlined above affects researchers and practitioners in AI.

2.6.1. Causal Attribution \neq Causal Explanation

An important concept is the relationship between cause attribution and explanation. Extracting a causal chain and displaying it to a person is not (necessarily) an explanation. While a person could use such a causal chain to obtain their own explanation, I argue that in many cases, this is merely causal attribution. In particular, for most models, it is not reasonable to expect a lay-user to be able to interpret a causal chain, no matter how it is presented. Much of the existing work in explainable AI literature is on the causal attribution part of explanation — something that, in many cases, is the easiest part of the problem because the causes are well understood, formalised, and accessible by the underlying models. In later sections, we will see more on the difference between attribution and explanation, why existing work in causal attribution is only part of the problem of explanation, and insights of how this work can be extended to produce more intuitive explanations.

2.6.2. Contrastive Explanation

Perhaps the most important point in this entire section is that explanation is contrastive (Section 2.3). Research indicates that people request only contrastive explanations, and that the cognitive burden of complete explanations is too great.

It could be argued that because models in AI operate at a level of abstraction that is considerably higher than real-world events, the causal chains are often smaller and less cognitively demanding, especially if they can be visualised. However, this argument misses a key point: it is not only the size of the causal chain that is important — people seem to be cognitively wired to process contrastive explanations, so one can argue that a layperson will find contrastive explanations more intuitive and more valuable.

This is both a challenge and an opportunity in AI. It is a challenge because eliciting a contrast case from a human observer may be difficult or even infeasible. Lipton [71] states that the obvious solution is that a non-contrastive question “*Why P?*” can be interpreted by default to “*Why P rather than not-P?*”. However, he then goes on to show that to answer “*Why P rather than not-P?*” is equivalent to providing all causes for *P*— something that is not so useful.

It is an opportunity because, as Lipton [71] argues, explaining a contrastive question is often easier than giving a full cause attribution because one only needs to understand what is different between the two cases, so one can provide a complete explanation without determining or even knowing all of the causes of the fact in question. This holds for computational explanation as well as human explanation.

Further, it can be beneficial in a more pragmatic way: it can help people construct mental models of systems to improve their understanding. For example, Kulesza et al. [63] describe the tricky problem of producing too much or too little information for explanatory debugging of machine learning models, but their results show that completeness (a person understanding an entire model) is more important than soundness, so too little information is more dangerous than too much. However, a contrast case provides a strong filter against this: if a person provides a foil, they are implicitly pointing towards the part of the model they do not understand. In Section 4.3, we will see research that outlines how people use contrasts to select explanations.

2.6.3. Explanatory Tasks and Levels of Explanation

Researchers and practitioners in explainable AI should understand and adopt a model of ‘levels of explanation’ — either one of those outlined above, or some other sensible model. The reason is clear: the answer that is provided to the why-question is strongly linked to the level at which the question is posed.

To illustrate, let’s take a couple of examples and apply them to Aristotle’s modes of explanation model outlined in Section 2.4. Consider our earlier arthropod classification algorithm from Section 1.4. At first glance, it may seem that such an algorithm resides at the *formal* level, so should offer explanations based on form. However, this would be erroneous, because that given categorisation algorithm has both efficient/mechanistic components, a reason for being implemented/executed (the *final* mode), and is implemented on hardware (the *final* mode). As such, there are explanations for its behaviour at all levels. It is hopefully not controversial to claim that *most* why-questions proposed by human observers about such an algorithm would indeed be at the formal level, posing questions such as “*Why is image J in group A instead of group B?*”, for which an answer could refer to the particular form of image and the groups A and B. However, in

our idealised dialogue, the question “*Why did you infer that the insect in image J had eight legs instead of six?*” asks a question about the underlying algorithm for counting legs, so the cause is at the efficient level; that is, it does not ask for what constitutes a spider in our model, but from where the inputs for that model came. Further, the final question about classifying the spider as an octopus refers to the final level, referring to the algorithms *function* or *goal*. Thus, causes in this algorithm occur at all four layers: (1) the material causes are at the hardware level to derive certain calculations; (2) the formal causes determine the classification itself; (3) the efficient causes determine such concepts as how features are detected; and (4) final causes determine why the algorithm was executed, or perhaps implemented at all.

As a second example, consider an algorithm for planning a robotic search and rescue mission after a disaster. In planning, programs are dynamically constructed, so different modes of cause/explanation are of interest compared to a classification algorithm. Causes still occur at the four levels: (1) the material level as before describes the hardware computation; (2) the formal level describes the underlying model passed to the planning tool; (3) the mechanistic level describe the particular planning algorithm employed; and (4) the final level describe the particular goal or intention of a plan. In such a system, the robot would likely have several goals to achieve; e.g. searching, taking pictures, supplying first-aid packages, returning to re-fuel, etc. As such, why-questions described at the final level (e.g. its goals) would be more common than in the classification algorithm example. However, the explaineer may also request questions related to the model, or why particular actions were taken rather than others, which may depend on the heuristics used or more likely, the particular property against which the algorithm is optimising (e.g. cost vs. time), and these require efficient/mechanistic explanations.

However, I am not arguing that we, as practitioners, must have explanatory agents capable of giving explanations at all of these levels — in particular, I pity any person that is tasked with designing a material explanation module for a machine learning or planning algorithm. This simply notes that these frameworks are useful for analysing the types of questions one may get. For an algorithm that categorises images, it seems that why-questions would *typically* be aimed at efficient and formal causes, because the purpose of the algorithm would in general be clear to its user, and that may be all one needs to provide in most contexts. In goal-based planning though, it is more likely that questions would related to the final and efficient levels, as we will see in Section 3.

In Sections 3 and 4, we will see work that demonstrates that for explanations at these different levels, people expect different types of explanation. Thus, it is important that we understand which types of questions refer to which levels in particular instances of technology, that different levels will be more useful/likely than others, and that, in research articles on interpretability, it is clear at which level we are aiming to provide explanations.

2.6.4. Model of Self

The work outlined in this section demonstrates that an intelligent must be able to reason about its own causal model. Consider our image classification example. When posed with the question “*Why is image J in group A instead of group B?*”, it is non-trivial, in my view, to attribute the cause by using the algorithm that generated the answer. A cleaner solution would be to have a more abstract symbolic model alongside this that records information such as when certain properties are detected and when

certain categorisations are made, which can be reasoned over. In other words, the agent requires a *model of self*⁴.

Hayes and Shah [40] note this in their work on explaining control policies, in which they select and annotate particular important state variables and actions that are relevant for explanation only. Langley et al. also hints at this idea, noting that “*An agent must represent content in a way that supports the explanations*” [66, p. 2], and an agent must have episodic memory to track many of its decisions, because it is more than just being able to understand the specific algorithm that made the decision — it must be explained in context.

Thus, to generate meaningful and useful explanations of behaviour, models based on the our understanding of explanation must sit alongside and work with the mechanisms that take the decisions under question.

2.6.5. Structure of Explanation

Related to the ‘model of self’ is the structure of explanation. Overton’s model of scientific explanation [105] defines what I believe to be a solid foundation for structure explanation in AI. In particular, this links with the notion of ‘model of self’. To provide an explanation along the chain outlined in Figure 3, one would need an explicit model of *each of these different categories* for the given system, as well as for the particular behaviour (or at least, to be able to generate them).

For example, the question from our dialogue in Section 1.4 “*How do you know that spiders have eight legs?*”, is a question referring not to the causal attribution in the classification algorithm itself, but is asking: “*How do you know this?*”, and thus is referring to how this was learnt — which, in this example, was learnt via another algorithm. In Overton’s model, this is not asking the question to explain the *Data* from the *Model*, but to explain the *Model* from the *Data* — in this case, the training data for the algorithm. This is more inline with scientific explanation. Such an approach requires an additional part of the ‘model of self’ that refers specifically to the learning, not the classification.

Why-questions can occur between any two categories in Overton’s model [105]. The question “*Why is image J labelled as a Spider instead of a Beetle?*” is asking to explain *Data* in terms of the *Model*, while the question “*How do you know that spiders have eight legs?*” is, as already noted, between *Model* and *Data*. However, there are questions between other layers: “*Why are eight-legged animals categorised as Spider instead of Beetle?*” is asking why a *Kind in general* is classified, not a particular instance. Researchers and practitioners should understand between which levels they believe explanations will be the most useful for particular applications.

A similar mapping can be made to the search and rescue planning example discussed above: an unmanned vehicle is an *Entity* of a particular *Kind*, which has a *Model* of behaviour defining how actions related to the beliefs and intentions of the agent. This *Model* is informed by relevant theories, such as folk psychology and even flight dynamics.

As with the levels of explanations, Overton’s model [105] or one similar to it seems necessary for researchers and practitioners in explainable AI to frame their thoughts and communicate their ideas.

⁴It was Darryn Reid who introduced me to this term.

3. Social Attribution — How Do People Explain Behaviour?

“Just as the contents of the nonsocial environment are interrelated by certain lawful connections, causal or otherwise, which define what can or will happen, we assume that there are connections of similar character between the contents of the social environment” – Heider [41, Chapter 2, pg. 21].

In this section, we outline work on *social attribution*, which defines how people attribute and (partly) explain behaviour of others. Such work is clearly relevant in many areas of artificial intelligence. However, research on social attribution laid the groundwork for much of the work outlined in Section 4, which looks at how people generate and evaluate events more generally, so we first start with research from social attribution that is most relevant to explainable AI. For a more detailed survey on this, see McClure [90] and Hilton [45].

3.1. Definitions

Social attribution is about *perception*. While the causes of behaviour can be described at a neurophysical level, and perhaps even lower levels, social attribution is concerned not with the real causes of human behaviour, but how other attribute or explain the behaviour of others. Heider [41] defines social attribution as *person perception*.

Intentions and *intentionality* were key to the work of Heider [41], and much of the recent work that has followed his — for example, Dennett [19], Malle [80], McClure [90], Boonzaier et al. [5], Kashima et al. [57]. An intention is a mental state of a person in which they form a commitment to carrying out some particular action or achieving some particular aim. Malle and Knobe [83] note that intentional behaviour therefore is always contrasted with *unintentional* behaviour, citing that laws of state, rules in sport, etc. all treat intentional actions different from unintentional actions. They note that, while intentionality can be considered an *objective* fact, it is also a *social* construct, in that people ascribe intentions to each other whether that intention is objective or not, and use these to socially interact.

Folk psychology, or commonsense psychology, is the attribution of human behaviour using ‘everyday’ terms such as beliefs, desires, intentions, emotions, and personality traits. This field of cognitive and social psychology recognises that, while such concepts may not truly cause human behaviour, these are the concepts that humans use to model and predict each others’ behaviours [80].

In the folk psychological model, actions consist of three parts: (1) the precondition of the action — that is, the circumstances under which it can be successfully executed, such as the capabilities of the actor or the constraints in the environment; (2) the action itself that can be undertaken; and (3) the effects of the action — that is, the changes that they bring about, either environmentally or socially.

Actions that are undertaken are typically explained by *goals* or intentions. In much of the work in social science, *goals* are equated with intentions. For our discussions, we define *goals* as being the end to which a mean contributes, while we define *intentions* as short-term goals that are adopted to achieve the end goals. The intentions have no utility themselves except to achieve positive utility goals. A *proximal* intention is a near-term intention that helps to achieve some further *distal* intention or goal. In the survey of existing literature, we will use the term used by the original authors, to ensure that they are interpreted as the authors expected.

3.2. Intentionality and Explanation

Heider [41] was the first person to experimentally try to identify how people attribute behaviour to others. In their now famous experiment from 1944, Heider and Simmel [42], showed a video containing animated shapes — a small triangle, a large triangle, and a small circle — moving around a screen⁵, and asked experiment participants to observe the video and then describe the behaviour of the shapes. Figure 4 shows a captured screenshot from this video in which the circle is opening a door to enter into a room. The participants’ responses described the behaviour anthropomorphically, assigning actions, intentions, emotions, and personality traits to the shapes. However, this experiment was not one on animation, but in social psychology. The aim of their 1944 experiment was to demonstrate that, as far as attributing causes to human behaviour, other people characterise deliberative behaviour using folk psychology.

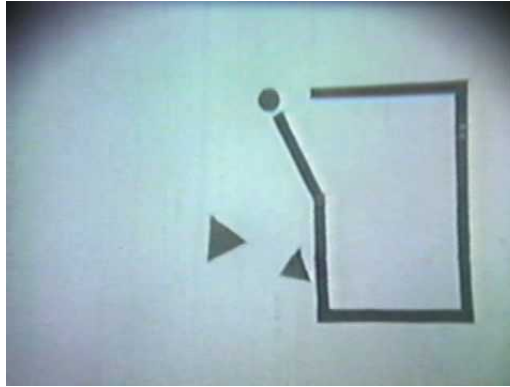


Figure 4: A screenshot of the video used in Heider and Simmel’s seminal study [42].

Heider [41] argued then that, the difference between *object perception* — describing causal behaviour of objects — and person perception was the intentions, or *motives*, of the person. He noted that behaviour in a social situation can have two types of causes: (1) *personal* (or *dispositional*) causality; and (2) *impersonal* causality, which can subsequently be influenced by *situational* factors, such as the environment. This interpretation lead to many researchers reflecting on the *person-situation* distinction and, in Malle’s view [82], incorrectly interpreting Heider’s work for decades.

Heider [41] contends that the key distinction between intentional action and non-intentional events is that intentional action demonstrates *equifinality*, which states that while the means to realise an intention may vary, the intention itself remains equa-final. Thus, if an actor should fail to achieve their intention, they will try other ways to achieve this intention, which differs from physical causality. Lombrozo [76] provides the example of Romeo and Juliet, noting that had a wall been placed between them, Romeo would have scaled the wall or knocked it down to reach his goal of seeing Juliet. However, iron filaments trying to get to a magnet would not display such equifinality — they would instead be simply blocked by the wall. Subsequent research confirms this distinction [19, 80, 90, 5, 57, 77].

⁵See the video here: <https://www.youtube.com/watch?v=VTNmLt7QX8E>.

Malle and Pearce [86] break the actions that people will explain into two dimensions: (1) *intentional vs. unintentional*; and (2) *observable vs. unobservable*; thus creating four different classifications (see Figure 5).

	Intentional	Unintentional
Observable	actions	mere behaviours
Unobservable	intentional thoughts	experiences

Figure 5: Malle’s classification of types of events, based on the dimensions of intentionality and observability [80, Chapter 3]

Malle and Pearce [86] performed experiments to confirm this model. As part of these experiments, participants were placed into a room with one other participants, and were left for 10 minutes to converse with each other to ‘get to know one another’, while their conversation was recorded. Malle and Pearce coded participants responses to questions with regards to observability and intentionality. Their results show that actors tend to explain unobservable events more than unobservable, which Malle and Pearce argue is because the actors are more *aware* of their own beliefs, desires, feelings, etc., than of their observable behaviours, such as facial expressions, gestures, postures, etc.). On the other hand, observers do the opposite for the inverse reason. Further, they showed that actors tend to explain unintentional behaviour more than intentional behaviour, again because (they believe) they are aware of their intentions, but not their ‘unplanned’ unintentional behaviour. Observers tend to find both intentional and unintentional behaviour difficult to explain, but will tend to find intentional behaviour more *relevant*.

3.3. Beliefs, Desires, Intentions, and Traits

Further to intentions, research suggest that other factors are important in attribution of social behaviour; in particular, beliefs, desires, and traits.

Kashima et al. [57] demonstrated that people use the folk psychological notions of belief, desire, and intention to understand, predict, and explain human action. In particular, they demonstrated that desires hold preference over beliefs, with beliefs being not explained if they are clear from the viewpoint of the explainee. They showed that people judge that explanations and behaviour ‘do not make sense’ when belief, desires, and intentions were inconsistent with each other. This early piece of work is one of the first to re-establish Heider’s theory of intentional behaviour in attribution [41].

However, it is the extensive body of work from Malle [79, 80, 81] that is the most seminal in this space.

3.3.1. Malle’s Conceptual Model for Social Attribution

Malle [80] proposes a model based on *Theory of Mind*, arguing that people attribute behaviour of others and themselves by assigning particular mental states that explain the behaviour. He offers six postulates (and sub-postulates) for the foundation of people’s folk explanation of behaviour, modelled in the scheme in Figure 6. He argues that these six postulates represent the assumptions and distinctions that people make when attributing behaviour to themselves and others:

1. People distinguish between intentional and unintentional behaviour.

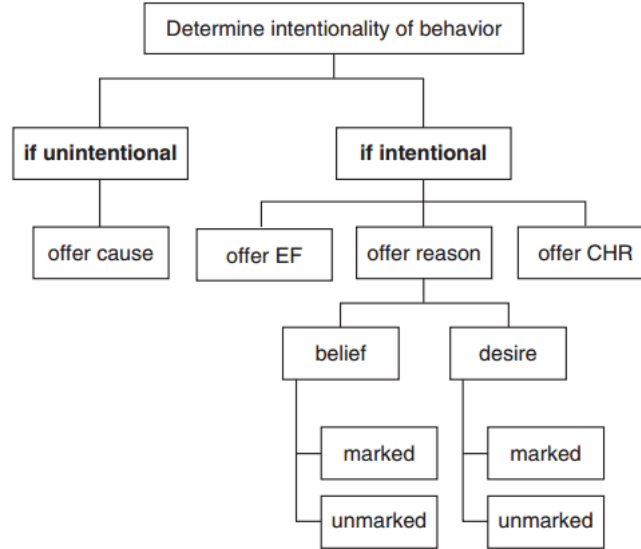


Figure 6: Malle’s conceptual framework for behaviour explanation; reproduced Malle [81, p. 87, Figure 3.3], adapted from Malle [80, p. 119, Figure 5.1]

2. For intentional behaviour, people use three modes of explanation based on the specific circumstances of the action:

- (a) *Reason explanations* are those explanations that link to the mental states (typically desires and beliefs, but also values) for the act, and the grounds on which they formed an intention.
- (b) *Causal History of Reason (CHR) explanations* are those explanations that use factors that “lay in the background” of an agents reasons (note, not the background of the action), but are not themselves reasons. Such factors can include unconscious motives, emotions, culture, personality, and the context. CHR explanations refer to causal factors that lead to reasons.

CHR explanations do not presuppose either subjectivity or rationality. This has three implications. First, they do not require the explainer to take the perspective of the explainee. Second, they can portray the actor as less rationale, by not offering a rational and intentional reason for the behaviour. Third, they allow the use of unconscious motives that the actor themselves would typically not use. Thus, *CHR explanations can make the agent look less rationale and in control than reason explanations*.

- (c) *Enabling factor explanations* are those explanations that explain not the intention of the actor, but instead explain how the intentional action achieved the outcome that it did. Thus, it assumes that the agent had an intention, and then refers to the factors that enabled the agent to successfully carry out the action, such as personal abilities or environmental properties. In essence, it relates to why preconditions of actions were enabled.

3. For unintentional behaviour, people offer just *causes*, such as physical, mechanistic, or habitual cases.

At the core of Malle’s framework is the intentionality of an act. For a behaviour to be considered intentional, the behaviour must be based on some *desire*, and a belief that the behaviour can be undertaken (both from a personal and situational perspective) and can achieve the desire. This forms the *intention*. If the agent has the ability and the awareness that they are performing the action, then the action is intentional.

Linguistically, people make a distinction between causes and reasons; for example, consider “*What were her reasons for choosing that book?*”, vs. “*What were his causes for falling over?*”. The use of “*his causes*” implies that the cause does not belong to the actor, but the reason does.

To give a reason explanation is to attribute *intentionality* to the action, and to identify the desires, beliefs, and valuing *in light of which* (subjectivity assumption) and *on the grounds of which* (rationality assumption) the agent acted. Thus, reasons imply intentionality, subjectivity, and rationality.

3.4. Individual vs. Group Behaviour

Susskind et al. [124] investigated how people ascribe causes to groups rather than individuals, focusing on traits. They provided experimental participants with a set of statements describing behaviours performed by individuals or groups, and were then asked to provide ratings of different descriptions of the people, such as their intelligence (a trait, or CHR in Malle’s framework), and were asked to judge the confidence of their judgements. Their results showed that as with individuals, participants freely assigned traits to groups, showing that groups are seen as agents themselves. However, they showed that when explaining an individual’s behaviour, the participants were able to produce explanations faster and more confidently than for groups, and that the traits that they assigned to individuals were judged to be less ‘extreme’ than those assigned to groups. In a second set of experiments, Susskind et al. showed that people expect more consistency in an individual’s behaviour compared to that of a group. When presented with a behaviour that *violated* the impression that participants had formed of individuals or groups, the participants were more likely to attribute the individual’s behaviour to causal mechanisms than the groups’ behaviour.

O’laughlin and Malle [103] further investigated people’s perception of group vs. individual behaviour, focusing on intentionality of explanation. In particular, they investigated the relative agency of groups that consist of ‘unrelated’ individuals acting independently (*aggregate groups*) compared to groups acting together (*jointly acting groups*).

Their study showed that people are more likely to offer explanations based on CHRs rather than intentions for aggregate groups, and more likely to offer explanations based on intentions than CHRs for jointly acting groups. For instance, to explain why all people in a department store came to that particular store, people are more likely offer a CHR explanation, such as that there was a sale on at the store that day. However, to answer the same question for why a group of friends are in the same department store, people are more likely to offer an explanation that the group wanted to spend the day together shopping – a desire. This seems to demonstrate that people feel that they cannot attribute intentional behaviour to the individuals in an aggregate group, so resort to more causal history explanations.

O’laughlin and Malle’s [103] finding that explanation of aggregate group behaviour is more likely to use CHRs is consistent with the earlier work from Kass and Leake [58], whose model of explanation explicitly divided *intentional* explanations from *social explanations*, which refers to explanations about human behaviour that is not intentionally driven (discussed in more detail in Section 2.4). These social explanations account for how people attribute deliberative behaviour to groups without referring to any form of intention.

An intriguing result from O’laughlin and Malle [103] is that while people attribute less intentionality to aggregate groups than to individuals, they attribute *more* intentionality to jointly acting groups than to individuals. O’laughlin and Malle account for this by reasoning that joint action is highly deliberative, so the group intention is more likely to have been explicitly agreed upon prior to acting, and the individuals within the group would be explicitly aware of this intention compared to the their own individual intentions.

3.5. Norms and Morals

Norms have been shown to hold a particular place in social attribution. Burguet and Hilton [9] (via Hilton [45]) showed that norms and abnormal behaviour are important in how people ascribe mental states to one another. For example, Hilton [45] notes that upon hearing the statement “*Ted admires Paul*”, people tend to attribute some trait to Paul as the object of the sentence, such as that Paul is charming and many people would admire him; and even that Ted does not admire many people. However, a counter-normative statement such as “*Ted admires the rapist*” triggers attributions instead to Ted, explained by the fact that it is non-normative to admire rapists, so Ted’s behaviour is distinctive to others, and is more likely to require an explanation. In Section 4, we will see more on the relationship between norms, abnormal behaviour, and attribution.

Uttich and Lombrozo [131] investigate the relationship of norms and the effect it has on attributing particular mental states, especially with regard to morals. They offer an interesting explanation of the *side-effect effect*, or the *Knobe effect* [61], which is the effect for people to attribute particular mental states (Theory of Mind) based on moral judgement. Knobe’s vignette from his seminal [61] paper is:

The vice-president of a company went to the chairman of the board and said, “We are thinking of starting a new program. It will help us increase profits, but it will also harm the environment”. The chairman of the board answered, “I don’t care at all about harming the environment. I just want to make as much profit as I can. Let’s start the new program.” They started the new program. Sure enough, the environment was harmed.

Knobe then produce a second vignette, which is exactly the same, but the side-effect of the program was in fact that the environment was *helped*. When participants were asked if the chairman had *intentionally* harmed the environment (first vignette), 82% of respondents replied yes. However, in the second vignette, only 23% thought that the chairman intentionally helped the environment.

Uttich and Lombrozo [131] hypothesise that the two existing camps aiming to explain this effect: the *Intuitive Moralist* and *Biased Scientist*, do not account for this. Instead, in their paper, they hypothesise that it is the fact the *norms* are violated that account

for this; that is, rather than moralist judgements influencing intentionality attribution, it is the more general relationship of conforming (or not) to norms (moral or not). In particular, behaviour that conforms to norms is less likely to change a person’s Theory of Mind (intention) of another person compared to behaviour that violates norms.

Samland and Waldmann [120] further investigate social attribution in the context of norms, looking at permissibility rather than obligation. They gave participants scenarios in which two actors combined to cause an outcome. For example, a department in which only administrative assistants are permitted to take pens from the stationary cupboard. One morning, Professor Smith (not permitted) and an assistant (permitted) each take a pen, and there are no pens remaining. Participants were tasked with rating how strongly each agent caused the outcome. Their results showed that participants rated the action of the non-permitted actor (e.g. Professor Smith) more than three times stronger than the other actor. However, if the outcome was positive instead of negative, such as an intern (not permitted) and a doctor (permitted) both signing off on a request for a drug for a patient, who subsequently recovers due to the double dose, participants rate the non-permitted behaviour only slightly stronger. As noted by Hilton [45, p. 54], these results indicate that in such settings, people seem to interpret the term *cause* as meaning “morally or institutionally responsible”.

In a follow-up study, Samland et al. [119] showed that children are not sensitive to norm violating behaviour in the same way that adults are. In particular, while both adults and children correlate cause and blame, children do not distinguish between cases in which the person was aware of the norm, while adults do.

3.6. Social Attribution and Artificial Intelligence

This section presents some ideas on how the work on social attribution outlined above affects researchers and practitioners in AI.

3.6.1. Folk Psychology

While the models and research results presented in this section pertain to the behaviour of humans, it is reasonably clear that these models have a place in explainable AI. Heider and Simmel’s seminal experiments from 1944 with moving shapes [42] (Section 3.2) demonstrate unequivocally that people attribute folk psychological concepts such as belief, desire, and intention, to artificial objects. Thus, we can conclude that in order to explain the behaviour of agents acting in an environment (physical or digital), particularly in a deliberative manner, building on a model based on the folk psychological concepts of beliefs, desires, intentions, etc., of the agent is valid.

This model is particularly promising because many knowledge-based models in deliberative AI either explicitly build on such folk psychological concepts, such as *belief-desire-intention* (BDI) models [113], or can be mapped quite easily to them; e.g. in classical-like AI planning, goals represent desires, intermediate/landmark states represent intentions, and the environment model represents beliefs [28].

In addition, the concepts and relationships between actions, preconditions, and proximal and distal intentions is similar to those in models such as BDI and planning, and as such, the work on the relationships between preconditions, outcomes, and competing goals, is useful in this area.

3.6.2. Malle's Models

Of all of the work outlined in this section, it is clear that Malle's model, culminating in his 2004 text book [80], is the most mature and complete model of social attribution to date. His three-layer model provides a solid foundation on which to build explanations of many deliberative systems, in particular, goal-based deliberation systems.

Malle's conceptual framework provides a suitable framework for characterising different aspects of causes for behaviour. It is clear that reason explanations will be useful for goal-based reasoners, as discussed in the case of BDI models and goal-directed AI planning. Further, as these types of models of deliberation often refer to preconditions of actions, enabling factor explanations can play a role in *how* questions and in counterfactual explanations. In Section 4, we will see further work on how to *select* explanations based on these concepts.

However, the causal history of reasons (CHR) explanations also have a part to play for deliberative agents. In human behaviour, they refer to personality traits and other unconscious motives. While anthropomorphic agents could clearly use CHRs to explain behaviour, such as emotion or personality, I argue that CHRs are also valid explanations for non-anthropomorphic agents. For example, consider an example of an AI planning system [28] that has produced a plan consisting of a sequence of actions, and a question posed asking why some action *a* was chosen instead of some action *b* as the final action in the sequence. A reason explanation would be used for cases in which action *a* achieved some particular part of the goal, while action *b* did not. An enabling factor representation would be used for cases in which action *a* was executable, but action *b* was not because its precondition was not satisfied. A CHR explanation would be used in the cases in which both achieved the goal and were executable, but that the cost of executing action *a* was less than action *b*. That is, the fact that the agent is optimising cost is a 'personality trait' of the agent that is invariant given the particular plan or goal.

Other types of planning systems may instead be risk averse, optimising to minimise risk or regret, or may be 'flexible' and try to help out their human collaborators as much as possible. These types of explanations are CHRs; even if they are not described as personality traits to the explainee. However, one must be careful to ensure these CHRs do not make their agent appear irrational — unless of course, that is the goal one is trying to achieve with the explanation process.

3.6.3. Collective Intelligence

The research into behaviour attribution of groups (Section 3.4) is important for those working in collective intelligence; areas such as in multi-agent planning [6], computational social choice [14], or argumentation [4]. Although this line of work appears to be much less explored than attributions of individual's behaviour, the findings from Kass and Leake [58], Susskind et al., and in particular O'Laughlin and Malle [103] that people assign intentions and beliefs to jointly-acting groups, and reasons to aggregate groups, indicates that the large body of work on attribution of individual behaviour could serve as a solid foundation for explanation of collective behaviour.

3.6.4. Norms and Morals

The work on norms and morals discussed in Section 3.5 demonstrates that normative behaviour, in particular, violation of such behaviour, has a large impact on the ascription of a Theory of Mind to actors. Clearly, for anthropomorphic agents, this work is

important, but as with CHRs, I argue here that it is important for more ‘traditional’ AI as well.

First, the link with morals is important for applications that elicit ethical or social concerns, such as defence, safety-critical applications, or judgements about people. Explanations or behaviour in general that violate norms may give the impression of ‘immoral machines’ — whatever that can mean — and thus, such norms need to be explicitly considered as part of explanation and interpretability.

Second, as discussed in Section 2.2, people mostly ask for explanations of events that they find unusual or abnormal [53, 49, 44], and violation of normative behaviour is one such abnormality [49]. Thus, it is clear that normative behaviour is important in interpretability — a statement that would not surprise those researchers and practitioners of normative artificial intelligence.

In Section 4, we will see that norms and violation of normal/normative behaviour is also important in the cognitive processes of people asking for, constructing, and evaluating explanations, and its impact on interpretability.

4. Cognitive Processes — How Do People Select and Evaluate Explanations?

“There are as many causes of x as there are explanations of x. Consider how the cause of death might have been set out by the physician as ‘multiple haemorrhage’, by the barrister as ‘negligence on the part of the driver’, by the carriage-builder as ‘a defect in the brakelock construction’, by a civic planner as ‘the presence of tall shrubbery at that turning’. None is more true than any of the others, but the particular context of the question makes some explanations more relevant than others.” – Hanson [38, p. 54].

Mill [98] is one of the earliest investigations of cause and explanation, and he argued that we make use of ‘statistical’ correlations to identify cause, which he called the *Method of Difference*. He argued that causal connection and explanation selection are essentially arbitrary, and scientifically/philosophically it is “wrong”, but offered several criteria that people seem to use, including things like unexpected conditions, precipitating causes, and variability. Such *covariation* models ideas were dominant in causal attribution, in particular, the work of Kelley [59]. However, many researchers started noting that the covariation models failed to explain many observations; for example, the people can identify causes between events from a single data point [95, 51]; and therefore, more recently, new theories have displaced them.

In this section, we look at these theories, in particular, we survey three types of cognitive processes used in explanation: (1) *causal connection*, which is process people use to identify the causes of events; (2) *explanation selection*, which is the process people use to select a small subset of the identified causes as *the* explanation; and (3) *explanation evaluation*, which is the processes that an explainee uses to evaluate the quality of an explanation.

4.1. Causal Connection, Explanation Selection, and Evaluation

Malle [80] presents a theory of explanation, which breaks the psychological processes used to offer explanations into two distinct groups, outlined in Figure 7:

1. *Information processes* — processes for devising and assembling explanations. The present section will present related work on this topic.
2. *Impression management processes* – processes for governing the social interaction of explanation. Section 5 will present related work on this topic.

Malle [80] further splits these two dimensions into two further dimensions, which refer to the tools for constructing and giving explanations, and the explainer’s perspective or knowledge about the explanation.

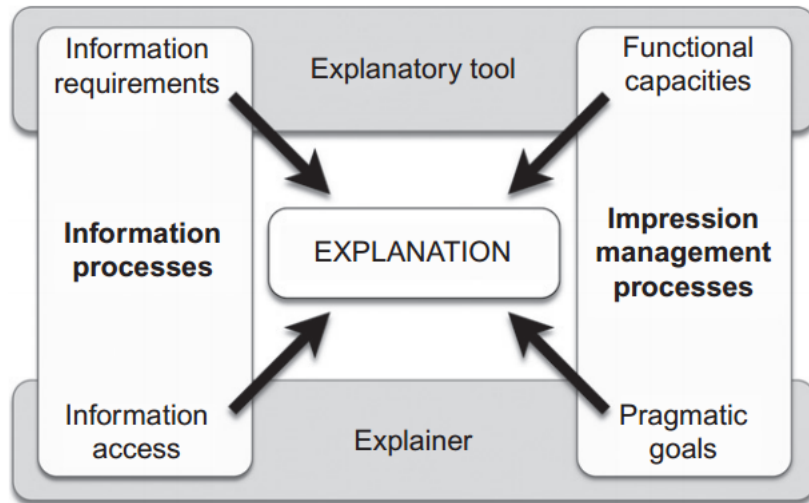


Figure 7: Malle’s process model for behaviour explanation; reproduced from Malle [82, p. 320, Figure 6.6]

Taking the two dimensions, there are four items:

1. *Information requirements* — what is required to give an adequate explanation; for example, one must know the causes of the explanandum, such as the desires and beliefs of an actor, or the mechanistic laws for a physical cause.
2. *Information access* — what information the explainer *has* to give the explanation, such as the causes, the desires, etc. Such information can be lacking; for example, the explainer does not know the intentions or beliefs of an actor in order to explain their behaviour.
3. *Pragmatic goals* — refers to the goal of the explanation, such as transferring knowledge to the explainee, making an actor look irrational, or generating trust with the explainee.
4. *Functional capacities* — each explanatory tool has functional capacities that constrain or dictate what goals can be achieved with that tool.

Malle et al. [85] argue that this theory accounts for apparent paradoxes observed in attribution theory, most specifically the actor-observer asymmetries, in which actors and

observers offer different explanations for the same action taken by an actor. He argues this is due to *information asymmetry*; e.g. an observer cannot access the intentions of an actor — the intentions must be inferred from the actor’s behaviour.

In this section, we first look specifically at processes related to the explainer: information access and pragmatic goals. When requested for an explanation, people typically do not have direct access to the causes, but infer them from observations and prior knowledge. Then, they select some of those causes as the explanation, based on the goal of the explanation. These two processes are known as *causal connection* (or *causal inference*), which is a processing of identifying the key causal connections the fact; and *explanation selection* (or *casual selection*), which is the processing of selecting a subset of those causes to provide as an explanation.

When explainees *receive* explanations, they go through the process of *explanation evaluation*, through which they determine whether the explanation is satisfactory or not. A primary criteria is that the explanation allows the explainee to understand the cause, however, other criteria exist as well in which people can consider some explanations to be better than others.

Background and Discounting. Hilton [49] discusses the complementary processes of *backgrounding* and *discounting* used by people in explanation. Discounting is when a hypothesis is deemed as less likely as a cause because additional context information is added to a competing hypothesis as part of causal connection. It is actually discounted as a cause to the event. Backgrounding involves pushing a possible cause to the background in the conversation of an explanation because it is not relevant to the goal, or new contextual information has been presented that make it no longer a good explanation (but still a cause). That is, while it is the cause of an event, it is not relevant to the explanation because the foil also has this cause.

As noted by Hilton [49], discounting occurs in the context of multiple possible causes — there are several possible causes and the person is trying to determine which causes the fact —, while backgrounding occurs in the context of multiple necessary events — a subset of necessary causes is selected as the explanation. Thus, discounting is part of causal connection, while backgrounding is part of explanation selection.

4.2. Causal Connection: Counterfactuals and Mutability

To determine the causes of anything other than a trivial event, it is not possible for a person to simulate back through all possible events and evaluate their counterfactual cases. Instead, people apply heuristics to select just some events to *mutate*. However, this process is not arbitrary. This section looks at several criteria that people assess the *mutability* of events; that is, the degree to which the event can be ‘undone’ to consider counterfactual cases. It shows that abnormality (including social abnormality), intention, time and controllability of events are key criteria.

4.2.1. Abnormality

Kahneman and Tversky [56] performed seminal work in this field, proposing the *simulation heuristic*. They hypothesised that when answering questions about past events, people perform a mental simulation of counterfactual cases. In particular, they show that abnormal events are mutable: they are the common events that people undo when judging causality. In their experiments, they asked people to identify primary causes

in causal chains using vignettes of a car accident causing the fatality of Mr. Jones, and which had multiple necessary causes (discussed in Section 4.3), including Mr. Jones going through a yellow light, and the teenager driver of the truck that hit Mr. Jones' car while under the influence of drugs. They used two vignettes: one in which Mr. Jones the car took an unusual route home to enjoy the view along the beach (the *route* version); and one in which he took the normal route home but left a bit early (the *time* version). Participants were asked to complete an 'if only' sentence that undid the fatal accident, imagining that they were a family member of Mr. Jones. Most participants in the route group undid the event in which Mr. Jones took the unusual route home more than those in the time version, while those in the time version undid the event of leaving early more often than those in the route version. That is, the participants tended to focus more on *abnormal* causes. In particular, Kahneman and Tversky note that people did not simply undo the event with the lowest prior probability in the scenario.

In their second study, Kahneman and Tversky [56] asked the participants to empathise with the family of the teenager driving the truck instead of with Mr. Jones, they found that people more often undid events of the teenage driver, rather Mr. Jones. Thus, the *perspective* or the *focus* is important in what types of events people undo.

More recently, Hilton et al. [46] showed that causal connection is relevant to the type of question that they pose. In particular, whether people are thinking *if only* some event had not occurred, or *but-for* some event. They gave participants scenarios similar to that of the accident scenario by Kahneman and Tversky [56], except that they added irrelevant events, such as the colour of Mr. Jones's car. Participants in both conditions discounted the colour of the car as causally irrelevant, but those in the *but-for* conditions generally accepted all other causes as relevant, in contrast to the *if-only* case, where participants mostly selected abnormal events. Asking people to put themselves into the perspective of a particular person, such as the family of Mr. Jones or the teenage boy, had no effect on *but-for* analysis, but did in the *if-only* cases. This demonstrates the importance of the question in being asked for causal connection.

4.2.2. Temporality

Miller and Gunasegaram [99] show that the *temporality* of events is important, in particular that people undo more recent events than more distal events. For instance, in one of their studies, they asked participants to play the role of a teacher selecting exam questions for a task. In one group, the *teacher-first* group, the participants were told that the students had not yet studied for their exam, while those in the another group, the *teacher-second* group, were told that the students had already studied for the exam. Those in the teacher-second group selected easier questions than those in the first, showing that participants perceived the degree of blame they would be given for hard questions depends on the temporal order of the tasks. This supports the hypothesis that earlier events are considered less mutable than later events.

4.2.3. Controllability and Intent

Giroto et al. [31] investigated mutability in causal chains with respect to *controllability*. They hypothesised that actions controllable by deliberative actors are more mutable than events that occur as a result of environmental effects. They provided participants with a vignette about Mr. Bianchi, who arrived late home from work to find his wife unconscious on the floor. His wife subsequently died. Four different events

caused Mr. Bianchi’s lateness: his decision to stop at a bar for a drink on the way home, plus three non-intentional causes, such as delays caused by abnormal traffic. Different questionnaires were given out with the events in different orders. When asked to undo events, participants overwhelmingly selected the intentional event as the one to undo, demonstrating that people mentally undo controllable events over uncontrollable events, irrelevant of the controllable events position in the sequence or whether the event was normal or abnormal. In another experiment, they varied whether the deliberative actions were *constrained* or *unconstrained*, in which an event is considered as constrained when they are somewhat enforced by other conditions; for example, Mr. Bianchi going to the bar (more controllable) vs. stopping due to an asthma attack (less controllable). The results of this experiment show that unconstrained actions are more mutable than constrained actions.

4.2.4. *Social Norms*

McCloy and Byrne [89] investigated the mutability of controllable events further, looking at the perceived appropriateness (or the socially normative perception) of the events. They presented a vignette similar to that of Girotto et al. [31], but with several controllable events, such as the main actor stopping to visit his parents, buy a newspaper, and stopping at a fast-food chain to get a burger. Participants were asked to provide ‘if only’ causes as before, as well as rate the ‘appropriateness’ of the behaviour. The results showed that participants were more likely to indicate inappropriate events as causal; e.g. stopping to buy a burger. In a second similar study, they showed that inappropriate events are traced through both normal and other exceptional events when identifying cause.

4.2.5. *Explanatory Modes*

As outlined in Section 2.4, philosophers and psychologists accept that different types of explanations exist; for example, Aristotle’s model: material, formal, efficient, and final. However, theories of causality have typically argued for only one type of cause, with the most prominent being dependence theories and transference theories.

Lombrozo [76] argues that the explanatory model both dependence theories and transference theories are at least *psychologically* real, even if only one (or neither) is the true theory. She hypothesises that people employ different modes of explanation for different modes of cognition, and thus both forms of explanation are valid: functional (final) explanations are better for phenomena that people consider have dependence relations, while mechanistic (efficient) explanations are better for physical phenomena.

Lombrozo [76] gave experimental participants scenarios in which the explanatory mode was manipulated and isolated using a mix of intentional and accidental/incidental human action, and in a second set of experiments, using biological traits that provide a particular function, or simply cause certain events incidentally. Participants were asked to evaluate different causal claims. The results of these experiments show that when events are interpreted in a functional manner, counterfactual dependence is important, but physical connections are not. However, when events are interpreted in a mechanistic manner, both counterfactual dependence and physical dependence are both deemed important. This implies that there is indeed a link between functional causation and dependence theories on the one hand, and between mechanistic explanation and transference theories on the other. Experiment participants also rated the functional

explanation stronger in the case that the causal dependence was intentional, as opposed to accidental.

Lombrozo [75] studied at the same issue of functional vs. mechanistic explanations for categorisation tasks specifically. Lombrozo presented participants with tasks similar to the following (text in square brackets added):

There is a kind of flower called a holing. Holings typically have brom compounds in their stems and they typically bend over as they grow. Scientists have discovered that having brom compounds in their stems is what usually causes holings to bend over as they grow [*mechanistic cause*]. By bending over, the holing’s pollen can brush against the fur of field mice, and spread to neighboring areas [*functional cause*].

Explanation prompt: *Why do holings typically bend over?*

They then gave participants a list of questions about flowers; for example: *Suppose a flower has brom compounds in its stem. How likely do you think it is that it bends over?*

Their results showed that participants who provided a mechanistic explanation were more likely to assign the objects with the proposed mechanistic feature than those who provided a functional explanation, and vice-versa. Their findings shows that giving explanations influences the importance of different features in the understanding of category membership, and that the importance of features in explanations can impact the categorisation of that feature.

4.2.6. Principles Connections in Formal Explanations

Prasada and Dillingham [110] and Prasada [109] discuss how people prioritise certain factors in the formal mode specifically; for instance, explaining why a particular arthropod is categorised as a beetle, such as in our running example. Prasada contends that “*Identifying something as an instance of a kind and explaining some of its properties in terms of its being the kind of thing it is are not two distinct activities, but a single cognitive activity.*” [109, p. 2]

Prasada and Dillingham [110] note that people represent relationships between the kinds of things, and the properties that they posses. This description conforms with Overton’s model of the structure of explanation [105] (see Section 2.6.5). Prasada and Dillingham’s experiments showed that people distinguish between two types of properties for a kind: *k-properties*, which are the properties of a thing that are due to its kind, which they call *principled connections*; and *t-properties*, which are the properties of a thing that are not due to its kind, which they call *factual connections*. Statistical correlations are examples of factual connections. For instance, a queen bee has a stinger and five legs because it is a bee (k-property), but the painted mark seen on almost all domesticated queen bees is because a bee keeper has marked it for ease of identification (t-property). K-properties have both principled and factual connections to their kind, whereas t-properties have mere factual connections. They note that k-properties have a *normative* aspect, in that it is expected that instances of kinds will have their k-properties, and when they do not, they are considered abnormal; for instance, a bee without a stinger.

In their experiments, the presented participants with explanations using different combinations of k-properties and t-properties to explain categorisations; for example,

“why is this a dog?” Their results showed that for formal modes, explanations involving k-properties were considered much better than explanations involving t-properties, and further, that using a thing’s kind to explain why it has a particular property was considered better for explaining k-properties than for explaining t-properties.

4.3. Explanation Selection

Similar to causal connection, people do not typically provide all causes for an event as an explanation. Instead, they *select* what they believe are the most relevant causes. Hilton [45] argues that explanation selection is used for cognitive reasons: causal chains are often too large to comprehend. He provides an example [45, p. 43, Figure 7] showing the causal chain for the story of the fatal car accident involving ‘Mr. Jones’ from Kahneman and Tversky [56]. For a simple story of a few paragraphs, the causal chain consists of over 20 events and 30 causes, all relevant to the accident. However, only a small amount of these are selected as explanations [129].

In this section, we overview key work that investigates the criteria people use for explanation selection. Perhaps unsurprisingly, the criteria for selection look similar to that of mutability, with temporality (proximal events preferred over distal events), abnormality, and intention being important, but also the features that are different between fact and foil.

4.3.1. Facts and Foils

As noted in Section 2, why-questions are contrastive between a fact and a foil. Research shows that the two contrasts are the primary way that people select explanations. In particular, to select an explanation from a set of causes, people look at the *difference* between the cases of the fact and foil.

Mackie [78] is one of the earliest to argue for explanation selection based on contrastive criteria, however, the first crisp definition of contrastive explanation seems to come from Hesslow [44]:

“This theory rests on two ideas. The first is that the effect or the explanandum, i.e. the event to be explained, should be construed, not as an object’s having a certain property, but as a difference between objects with regard to a certain property. The second idea is that selection and weighting of causes is determined by explanatory relevance.” [Emphasis from the original source]
— Hesslow [44, p. 24].

Hesslow [44] argues that criteria for selecting explanations are clearly not arbitrary, because people seem to select explanations in similar ways to each other. He defines an explanan as a relation containing an object *a* (the *fact* in our terminology), a set of comparison objects *R*, called the *reference class* (the *foils*), and a property *E*, which *a* has but the objects in reference class *R* does not. For example, *a* = Spider, *R* = Beetle, and *E* = eight legs. Hesslow argues that the contrast between the fact and foil is the primary criteria for explanation selection, and that the explanation with the highest *explanatory power* should be the one that highlights the greatest number of *differences* in the attributes between the target and reference objects.

Lipton [71], building on earlier work in philosophy from Lewis [69], derived similar thoughts to Hesslow [44], without seeming to be aware of his work. He proposed a definition of contrastive explanation based on what he calls the *Difference Condition*:

“To explain why P rather than Q, we must cite a causal difference between P and not-Q, consisting of a cause of P and the absence of a corresponding event in the history of not-Q.” – Lipton [71, p. 256].

Returning to our arthropod example, for the why-question between image *J* categorised as a fly and image *K* a beetle, image *J* having six legs is correctly determined to have no explanatory relevance, because it does not cause *K* to be categorised as a beetle instead of a fly. Instead, the explanation would cite some other cause, which according to Table 1, would be that the arthropod in image *J* has five eyes, consistent with a fly, while the one in image *K* has two, consistent with a beetle.

From an experimental perspective, Hilton and Slugoski [53] were the first researchers to both identify the limitations of covariation, and instead propose that contrastive explanation is best described as the differences between the two events (discussed further in Section 4.3.2). More recent research in cognitive science Rehder [115, 116] supports the theory that people perform causal inference, explanation, and generalisation based on contrastive cases.

4.3.2. Abnormality

Related to the idea of contrastive explanation, Hilton and Slugoski [53] propose the *abnormal conditions model*, based on observations from legal theorists Hart and Honoré [39]. Hilton and Slugoski argue that *abnormal* events play a key role in causal explanation. They argue that, while statistical notions of co-variance are not the only method employed in everyday explanations, the basic idea that people select unusual events to explain is valid. Their theory states that explainers use their perceived background knowledge with explainees to select those conditions that are considered *abnormal*. They give the example of asking why the Challenger shuttle exploded in 1986 (rather than not exploding, or perhaps why most other shuttles do not explode). The explanation that it exploded “because of faulty seals” seems like a better explanation than “there was oxygen in the atmosphere”. The abnormal conditions model accounts for this nicely by noting that an explainer will reason that oxygen is present in the atmosphere when all shuttles launch, so this is not an abnormal condition. On the other hand, most shuttles do not have faulty seals, so this contributing factor was a necessary yet abnormal event in the Challenger disaster.

The abnormal conditions model has been backed up by subsequent experimental studies, such as those by McClure and Hilton [93], McClure et al. [94], and Hilton et al. [52], and more recently, Samland and Waldmann [120], who show that a variety of non-statistical measures are valid foils.

4.3.3. Intentionality and Functionality

Other features of causal chains have been demonstrated to be more important than abnormality.

Hilton et al. [52] investigate the claim from legal theorists Hart and Honoré [39] that intentional action takes priority of non-intentional action in opportunity chains. Their perspective builds on the abnormal conditions model, noting that there are two important contrasts in explanation selection: (1) normal vs. abnormal; and (2) intentional vs. non-intentional. They argue further that causes will be “traced through” a proximal (more recent) abnormal condition if there is a more distal (less recent) event that is intentional.

For example, to explain why someone died, one would explain that the poison they ingested as part of a meal was the cause of death; but if the poison as shown to have been deliberately placed in an attempt to murder the victim, the intention of someone to murder the victim receives priority. In their experiments, they gave participants different opportunity chains in which a proximal abnormal cause was an intentional human action, an unintentional human action, or a natural event, depending on the condition to which they were assigned. For example, a cause of an accident was ice on the road, which was enabled by either someone deliberately spraying the road, someone unintentionally placing water on the road, or water from a storm. Participants were asked to rate the explanations. Their results showed that: (1) participants rated intentional action as a better explanation than the other two causes, and non-intentional action better than natural cases; and (2) in opportunity chains, there is little preference for proximal over distal events if two events are of the same type (e.g. both are natural events) — both are seen as necessary.

Hilton et al. [51] extend this work to unfolding causal chains in causal explanation, in particular, they investigate transitive unfolding causal chains. Running a similar set of experiments as Hilton et al. [52], they had similar findings with regards to intentional actions vs. other events: participants traced through abnormal events to the more distal intentional action. However, in unfolding chains — as opposed to opportunistic chains — the result was different: more distal intentional actions are considered a better explanation than a more proximal causes that are non-intentional.

Lombrozo [76] argues further that this holds for *functional* explanations in general; not just intentional action. For instance, citing the functional reason that an object exists is preferred to mechanistic explanations.

4.3.4. *Necessity*

Lipton [71] argues that *necessary* causes are preferred to sufficient causes. For example, consider mutations in the DNA of a particular species of beetle that cause its wings to grow longer than normal when kept in certain temperatures. Now, consider that there is two such mutations, M_1 and M_2 , and either is sufficient to cause the mutation. To contrast with a beetle whose wings would not change, the explanation of temperature is preferred to either of the mutations M_1 or M_2 , because neither M_1 nor M_2 are individually necessary for the observed event; merely that *either* M_1 or M_2 . In contrast, the temperature is necessary, and is preferred, even if we know that the cause was M_1 .

4.3.5. *Responsibility*

The notions of *responsibility* and *blame* are relevant to causal selection, in that an event considered the more responsible for an outcome is likely to be judged as a better explanation than other causes. In fact, it relates closely to necessity, as responsibility aims to place a measure of degrees of necessity of causes. An event that is fully responsible outcome for an event is a necessary cause.

Chockler and Halpern [16] modified the structural equation model proposed by Halpern and Pearl [35] (see Section 2.1.1) to define responsibility of an outcome. Informally, they define the responsibility of cause C to event E under a situation based on the minimal number of changes required to the situation to make event E no longer occur. If N is the minimal number of changes required, then the responsibility of C causes E is $\frac{1}{N+1}$.

If $N = 0$, then C is fully responsible. Thus, one can see that an event that is considered more responsible than another requires less changes to prevent E than the other.

While several different cognitive models of responsibility attribution have been proposed (c.f. [50, 65]), I focus on the model of Chockler and Halpern [16] because, as far I am aware, experimental evaluation of the model shows it to be stronger than existing models [27], and because it is a formal model that is more readily adopted in artificial intelligence.

The structural model approach defines the responsibility of *events*, rather than individuals or groups, but one can see that it can be used in group models as well. Gerstenberg and Lagnado [27] show that the model has strong predictive power at attributing responsibility to individuals in groups. They ran a set of experiments in which participants played a simple game in teams in which each individual was asked to count the number of triangles in an image, and teams won or lost depending on how accurate their *collective* counts were. After the game, participants rated the responsibility of each player to the outcome. Their results showed that the modified structural equation model Chockler and Halpern [16] was more accurate at predicting participants outcomes than simple counterfactual model and the so-called *Matching Model*, in which the responsibility is defined as the degree of deviation to the outcome; in the triangle counting game, this would be how far of the individual was to the actual number of triangles.

4.3.6. Preconditions, Failure, and Intentions

An early study into explanation selection in cases of more than one cause was undertaken by Leddo et al. [67]. They conducted studies asking people to rate the probability of different factors as causes of events. As predicted by the intention/goal-based theory, goals were considered better explanations than relevant preconditions. However, people also rated conjunctions of preconditions and goals as better explanations of why the event occurred. For example, for the action “*Fred went to the restaurant*”, participants rated explanations such as “*Fred was hungry*” more likely than “*Fred had money in his pocket*”, but further “*Fred was hungry and had money in his pocket*” as an even more likely explanation, despite the fact the cause itself is less likely (conjoining the two probabilities). This is consistent with the well-known *conjunction fallacy* [130], which shows that people sometimes estimate the probability of the conjunction of two facts higher than either of the individual fact if those two facts are representative of prior beliefs.

However, Leddo et al. [67] further showed that for *failed or uncompleted* actions, just one cause (goal or precondition) was considered a better explanation, indicating that failed actions are explained differently. This is consistent with physical causality explanations [75]. Leddo et al. argue that to explain an action, people combine their knowledge of the particular situation with a more general understanding about causal relations. Lombrozo [76] argues similarly that this is because failed actions are not goal-directed, because people do not intend to fail. Thus, people like prefer *mechanistic* explanations for failed actions, rather than explanations that cite intentions.

McClure and Hilton [91] and McClure et al. [92] found that people tend to assign a higher probability of conjoined goal and precondition for a successful action, even though they prefer the goal as the best explanation, *except* in extreme/unlikely situations; that is, when the precondition is unlikely to be true. They argue that is largely due to the (lack of) *controllability* of unlikely actions. That is, extreme/unlikely events are judged to be harder to control, and thus actors would be less likely to intentionally select that

action *unless* the unlikely opportunity presented itself. However, for normal and expected actions, participants preferred the goal alone as an explanation instead of the goal and precondition.

In a follow-up study, McClure and Hilton [93] looked at explanations of obstructed vs. unobstructed events, in which an event is obstructed by its precondition being false; for example, “*Fred wanted a coffee, but did not have enough money to buy one*” as an explanation for why Fred failed to get a coffee. They showed that while goals are important to both, for obstructed events, the precondition becomes more important than for unobstructed events.

4.4. Explanation Evaluation

In this section, we look at work that has investigated the criteria that people use to evaluate explanations given to them. The most important of these are: probability, simplicity, generalise, and coherence with prior beliefs.

4.4.1. Coherence, Simplicity, and Generality

Thagard [128] argues that coherence is a primary criteria for explanation. He proposes the *Theory for Explanatory Coherence*, which specifies seven principles of how explanations relate to prior belief. He argues that these principles are foundational principles that explanations must observe to be acceptable. They capture properties such as if some set of properties P explain some other property Q , then all properties in P must be coherent with Q ; that is, people will be more likely to accept explanations if they are consistent with their prior beliefs. Further, he contends that all things being equal, simpler explanations — those that cite fewer causes — and more general explanations — those that explain more events —, are better explanations. The model has been demonstrated to align with how humans make judgements on explanations [112].

Read and Marcus-Newhall [114] tested the hypotheses from Thagard’s theory of explanatory coherence [128] that people prefer simpler and more general explanations. They constructed a set of experiments in which participants were asked to rate the probability and the ‘quality’ of the explanations with different numbers of causes. Participants were given stories containing several events to be explained, and several different explanations. For example, one story was about Cheryl, who is suffering from three medical problems: (1) weight gain; (2) fatigue; and (3) nausea. Different participant groups were given one of three types of explanations: (1) *narrow*: one of Cheryl having stopped exercising (weight gain), has mononucleosis (explains fatigue), or a stomach virus (explains nausea); (2) *broad*: Cheryl is pregnant (explains all three); or (3) *conjunctive*: all three from item 1 as the same time. As predicted, participants preferred simple explanations (pregnancy) with less causes than more complex ones (all three conjunctions), and participants preferred explanations that explained more events.

4.4.2. Truth and Probability

An initial criteria one can propose is that explanations should be true or likely. However, Hilton [49] shows that the most likely or ‘true’ cause is not necessarily the *best* explanation. Truth conditions⁶ are a necessary but not sufficient criteria for the genera-

⁶We use the term *truth condition* to refer to facts that are either true or considered likely by the explaineer.

tion of explanations. While a true or likely cause is one attribute of a good explanation, tacitly implying that the most probable cause is always the best explanation is incorrect. As an example, consider again the explosion of the Challenger shuttle (Section 4.3.2), in which a faulty seal was argued to be a better explanation than oxygen in the atmosphere. This is despite the fact the the ‘seal’ explanation is a likely but not known cause, while the ‘oxygen’ explanation is a known cause. Hilton argues that this is because the fact that there is oxygen in the atmosphere is *presupposed*; that is, the explainer assumes that the explainee already knows this.

McClure [90] also challenges the idea of probability as a criteria for explanations. Their studies found that people tend not to judge the quality of explanations around their probability, but instead around their so-called *pragmatic influences* of causal behaviour. That is, people judge explanations on their usefulness, relevance, etc., including via Grice’s maxims of conversation [33] (see Section 5.1.1 for a more detailed discussion of this). This is supported by experiments such as Read and Marcus-Newhall [114] cited above, and the work from Tversky and Kahneman [130] on the conjunction fallacy.

Lombrozo [74] notes that the experiments on generality and simplicity performed by Read and Marcus-Newhall [114] cannot rule out that participants selected simple explanations because they did not have probability or frequency information for events. Lombrozo argues that if participants assumed that the events of stopping exercising, having mononucleosis, having a stomach virus, and being pregnant are all equally likely, then the probability of the conjunction of any three is much more unlikely than any one combined. To counter this, she investigated the influence that probability has on explanation evaluation, in particular, when simpler explanations are less probable than more complex ones. Based on a similar experimental setup to that of Read and Marcus-Newhall [114], Lombrozo presented experimental participants with information about a patient with several symptoms that could be explained by one cause or two separate causes. In some setups, base rate information about each disease was provided, in which the conjunction of the two separate causes was more likely than the single (simpler) cause. Without base-rate information, participants selected the most simple (less likely) explanations, but when base-rate information was included, the difference was less pronounced. However, the likelihood of the conjunctive scenario had to be *significantly more likely for it to be chosen*. Lombrozo’s final experiment showed that this effect was reduced again if participants were *explicitly* provided with the joint probability of the two events, rather than in earlier experiments in which they were provided separately.

Preston and Epley [111] show that the value that people assign to their own beliefs – both in terms of probability and personal relevance – correspond with the *explanatory power* of those beliefs. Participants were each given a particular ‘belief’ that is generally accepted by psychologists, but mostly unknown in the general public, and were then allocated to three conditions: (1) the *applications* condition, who were asked to list observations that the belief could explain; (2) the *explanations* condition, who were asked to list observations that could explain the belief (the inverse to the previous condition); and (3) a control condition who did neither. Participants were then asked to consider the probability of that belief being true, and to assign their perceived *value* of the belief to themselves and society in general. Their results show that people in the applications and explanations condition both assigned a higher probability to the belief being true, demonstrating that if people link beliefs to certain situations, the perceived probability increased. However, for value, the results were different: those in the applications condi-

tion assigned a higher value than the other two conditions, and those in the explanations condition assigned a lower value than the other two conditions. This indicates that people assign higher values to beliefs that explain observations, but a lower value to beliefs that can be explained by *other* observations.

4.4.3. Goals and Explanatory Mode

Vasilyeva et al. [133] show that the goal of explainer is key in how the evaluated explanations, in particular, in relation to the *mode* of explanation used (i.e. material, formal, efficient, final). In their experiments, they gave participants different tasks with varying goals. For instance, some participants were asked to assess the causes behind some organisms having certain traits (efficient), others were asked to categorise organisms into groups (formal), and the third group were asked for what reason organisms would have those traits (functional). They provided explanations using different modes for parts of the tasks and then asked participants to rate the ‘goodness’ of an explanation provided to them. Their results showed that the goals not only shifted the focus of the questions asked by participants, but also that participants preferred modes of explanation that were more congruent with the goal of their task. This is further evidence that being clear about the question being asked is important in explanation.

4.5. Cognitive Processes and Artificial Intelligence

This section presents some ideas on how the work on the cognitive processes of explanation affects researchers and practitioners in AI.

4.5.1. Mutability and Computation

Section 4.2 studies the heuristics that people use to discount some events over others during mental simulation of causes. This is relevant to some areas of explainable AI because, in the same way that people apply these heuristics to more efficiently search through a causal chain, so too can these heuristics be used to more efficiently find causes, while still identifying causes that a human explainee would expect.

The notions of causal temporality and responsibility would be reasonably straightforward to capture in many models, however, if one can capture concepts such as abnormality, responsibility intentional, or controllability in models, this provides further opportunities.

4.5.2. Abnormality

Abnormality clearly plays a role in explanation and interpretability. For explanation, it serves as a trigger for explanation, and is a useful criteria for explanation selection. For interpretability, it is clear that ‘normal’ behaviour will, on aggregate, be judged more explainable than abnormal behaviour.

Abnormality is a key criteria for explanation selection, and as such, the ability to identify abnormal events in causal chains could improve the explanations that can be supplied by an explanatory agent. While for some models, such as those used for probabilistic reasoning, identifying abnormal events would be straightforward, and for others, such as normative systems, they are ‘built in’, for other types of models, identifying abnormal events could prove difficult.

One important note to make is regarding abnormality and its application to “*non-contrastive*” why-questions. As noted in the previous subsection, interpreting the question “*Why P?*” as “*Why P rather than not-P?*” offers no real benefit: they are both equivalent to providing a complete cause of P . However, it could be that the concept of normality can be used to mitigate this problem. That is, in the case of “*Why P?*”, we can interpret this as “*Why P rather than the normal case Q?*” [48]. For example, consider the application of assessing the risk of glaucoma [11]. Instead of asking why they were given a positive diagnosis rather than a negative diagnosis, the explanatory again could provide one or more *default* foils, which would be ‘stereotypical’ examples of people who were not diagnosed and whose symptoms were more regular with respect to the general population. Then, the question becomes why was the person diagnosed with glaucoma compared to these default stereotypical cases without glaucoma.

4.5.3. *Intentionality and Functionality*

The work discussed in Section 4.3.3 demonstrates the importance of intentionality and functionality in selecting explanations. As discussed in Section 3.6.1, these concepts are highly relevant to deliberative AI systems, in which concepts such as goals and intentions are first-class citizens. However, the importance of this to explanation selection rather than social attribution must be drawn out. In social attribution, folk psychological concepts such as intentions are attributed to agents to identify causes and explanations, while in this section, intentions are used as part of the cognitive process of *selecting* explanations from a causal chain. Thus, even for a non-deliberative system, labelling cause types as intention could be useful. For instance, consider a predictive model in which some features represent that an intentional event has occurred. These features could be selected or prioritised in explanations.

4.5.4. *Perspectives and Controllability*

The finding from Kahneman and Tversky [56] that perspectives change the events people mutate, discussed in Section 4.2, is important in multi-agent contexts. From an explanation in AI perspective, this implies that when generating explanations to explain a particular agent’s decisions or behaviour, the explanatory agent should focus on undoing actions of that particular agent, rather than others. This is also consistent with the research on controllability discussed in Section 4.2, in that, from the perspective of the agent in question, they can only control their own actions.

In interpretability, the impact of this work is also clear: in generating explainable behaviour, with all others things being equal, agents could select actions that lead to future actions being more constrained, as the subsequent actions are less likely to have counterfactuals undone by the observer.

4.5.5. *Evaluation of Explanations*

The importance of the research outlined in Section 4.4 is clear: likelihood is not everything. While likely causes are part of good explanations, they do not strongly correlate with explanations that people find useful. The work outlined in this section provides three criteria that are at least as equally important: simplicity, generality, and coherence.

For explanation, if the goal of an explanatory agent is to provide the most likely causes of an event, then these three criteria can be used to prioritise among the most

likely events. However, if the goal of an explanatory agent is to generate trust between itself and its human observers, these criteria should be considered as first-class criteria in explanation generation beside or even above likelihood. For example, providing simpler explanations that increase the likelihood that the observer both *understands* and *accepts* the explanation may increase trust better than giving more likely explanations.

For interpretability, similarly, these three criteria can form part of decision-making algorithms; for example, a deliberative agent may opt to select an action that is slightly less likely to achieve its goal, if the action helps towards other goals that the observer knows about, and has a smaller number of causes to refer to.

5. Social Explanation — How Do People Communicate Explanations?

*“Causal explanation is first and foremost a form of social interaction. One speaks of giving causal explanations, but not attributions, perceptions, comprehensions, categorizations, or memories. The verb to explain is a three-place predicate: **Someone** explains **something** to **someone**. Causal explanation takes the form of conversation and is thus subject to the rules of conversation.”* [Emphasis original] — Hilton [48].

This final section looks at the communication problem in explanation — something that has been studied little in explainable AI so far. The work outlined in this section asserts that the explanation process does not stop at just selecting an explanation, but considers that an explanation is an interaction between two roles: explainer and explainee (perhaps the same person/agent playing both roles), and that there are certain ‘rules’ that govern this interaction.

5.1. Explanation as Conversation

Hilton [48] presents the most seminal article on the social aspects of conversation, proposing a *conversational model of explanation* based on foundational work undertaken by both himself and others. The primary argument of Hilton is that explanation is a *conversation*, and this is how it differs from causal attribution. He argues that there are two stages: the *diagnosis* of causality in which the explainer determines why an action/event occurred; and the *explanation* is the social process of conveying this to someone. The problem is then to “*resolve a puzzle in the explainee’s mind about why the event happened by closing a gap in his or her knowledge*” [48, p. 66].

The conversational model argues that good social explanations must be *relevant*. This means that they must answer the *why* question that is asked — merely identifying causes does not provide good explanations, because many of the causes will not be relevant to the questions; or worst still, if the “most probable” causes are selected to present to the explainee, they will not be relevant to the question asked. The information that is communicated between explainer and explainee should conform to the general rules of *cooperative* conversation [33], including being relevant to the explainee themselves, and what they already know.

Hilton [48] terms the second stage *explanation presentation*, and argues that when an explainer presents an explanation to an explainee, they are engaged in a conversation. As such, they tend to follow basic rules of conversation, which Hilton argues are captured by *Grice’s maxims of conversation* [33]: (a) quality; (b) quantity; (c) relation; and (d)

manner. Coarsely, these respectively mean: only say what you believe; only say as much as is necessary; only say what is relevant; and say it in a nice way.

These maxims imply that the shared knowledge between explainer and explainee are *presuppositions* of the explanations, and the other factors are the causes that should be explained; in short, the explainer should not explain any causes they think the explainee already knows.

Previous sections have presented the relevant literature about causal connection (Sections 3 and 4) and explanation selection (Sections 4). In the remainder of this subsection, we describe Grice’s model and present related research that analyses how people select explanations relative to subjective (or social) viewpoints, and present work that supports Hilton’s conversational model of explanation [48].

5.1.1. Logic and Conversation

Grice’s maxims [33] (or the *Gricean maxims*) are a model for how people engage in cooperative conversation. Grice observes that conversational statements do not occur in isolation: they are often linked together, forming a cooperative effort to achieve some goal of information exchange or some social goal, such as social bonding. He notes then that a general principle that one should adhere to in conversation is the *cooperative principle*: “*Make your conversational contribution as much as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged*” [33, p. 45].

For this, Grice [33] distinguishes four categories of maxims that would help to achieve the cooperative principle:

1. *Quality*: Make sure that the information is of high quality – try to make your contribution one that is true. This contains two maxims: (a) do not say things that you believe to be false; and (b) do not say things for which you do not have sufficient evidence.
2. *Quantity*: Provide the right quantity of information. This contains two maxims: (a) make your contribution as informative as is required; and (b) do not make it more informative than is required.
3. *Relation*: Only provide information that is related to the conversation. This consists of a single maxim: (a) Be relevant. This maxim can be interpreted as a strategy for achieving the maxim of quantity.
4. *Manner*: Relating to *how* one provides information, rather than what is provided. This consists of the ‘supermaxim’ of ‘Be perspicuous’, but according to Grice, is broken into ‘various’ maxims such as: (a) avoid obscurity of expression; (b) avoid ambiguity; (c) be brief (avoid unnecessary prolixity); and (d) be orderly.

Grice [33] argues that for cooperative conversation, one should obey these maxims, and that people learn such maxims as part of their life experience. He further links these maxims to *implicature*, and shows that it is possible to violate some maxims while still being cooperative, in order to either not violate one of the other maxims, or to achieve some particular goal, such as to *implicate* something else without saying it. Irony and metaphors are examples of violating the quality maxims, but other examples, such as:

Person A: “*What did you think of the food they served?*”; Person B: “*Well, it was certainly healthy*”, violates the maxim of manner, but is implying perhaps that Person B did not enjoy the food, without them actually saying so.

Following from the claim that explanations are conversations, Hilton [48] argues that explanations should follow these maxims. The quality and quantity categories present logical characterisations of the explanations themselves, while the relation and manner categories define how they explanations should be given.

5.1.2. *Relation & Relevance in Explanation Selection*

Of particular interest here is research to support these Gricean maxims; in particular, the related maxims of *quantity* and *relevance*, which together state that the speaker should only say what is necessary and relevant. In social explanation, research has shown that people select explanations to adhere to these maxims by considering the particular question being asked by the explainee, but also by giving explanations that the explainee does not already accept as being true. To quote Hesslow [44, p. 30]:

“What are being selected are essentially questions, and the causal selection that follows from this is determined by the straightforward criterion of explanatory relevance.”

In Section 4.3.1, we saw evidence to suggest that the difference between the fact and foil for contrastive why-question are the relevant causes for explanation. In this section, we review work on the social aspects of explanation selection and evaluation.

Epistemic Relevance. Slugoski et al. [122] present evidence of Gricean maxims in explanation, and of support for the idea of explanation as conversation. They argue that the form of explanation must take into account its function as an answer to a specified why-question, and that this should take part within a conversational framework, including the context of the explainee. They gave experimental participants information in the form of a police report about an individual named George who had been charged with assault after a school fight. This information contained information about George himself, and about the circumstances of the fight. Participants were then paired with another ‘participant’ (played by the researcher), were told that the other participant had either: (a) information about George; (2) the circumstances of the fight; or (c) neither; and were asked to answer why George had assaulted the other person. The results showed participants provided explanations that are tailored to their expectations of what the hearer already knows, selecting single causes based on abnormal factors of which they believe the explainee is unaware; and that participants *change* their explanations of the same event when presenting to explainees with differing background knowledge.

Jaspars and Hilton [55] and Hilton [49] both argue that such results demonstrate that, as well as being true or likely, a good explanation must be relevant to both the question and to the *mental model* of the explainee. Byrne [10] offers a similar argument in her computational model of explanation selection, noting that humans are model-based, not proof-based, so explanations must be relevant to a model.

Halpern and Pearl [36] present an elegant formal model of explanation selection based on epistemic relevance. This model extends their work on structural causal models [36], discussed in Section 2.1.1. They define an explanation as a fact that, if found to be true, would constitute an actual cause of a specific event.

Recall from Section 2.1.1 structural causal models [35] contain variables and functions between these variables. A *situation* is a unique assignment from variables to values. Halpern and Pearl [36] then define an *epistemic state* as a set of situations, one for each possible situation that the explainee believes is possible. Explaining the causes of an event then becomes providing the values for those variables that ‘shrinks’ the epistemic state down to a size for which the cause of the event can be uniquely identified. They then further show how to provide explanations that describe the structural model itself, rather than just the values of variables, and how to reason when provided with probability distributions over events. Given a probabilistic model, Halpern and Pearl formally define the *explanatory power* of partial explanations. Informally, this states that explanation C_1 has more explanatory power explanation C_2 for explanandum E if and only if providing C_1 to the explainee increases the prior probability of E being true more than providing C_2 does.

Dodd and Bradshaw [20] demonstrates that the perceived intention of a speaker is important in implicature. Just as leading questions in eyewitness reports can have an effect on the judgement of the eyewitness, so to it can affect explanation. They showed that the meaning and presuppositions that people infer from conversational implicatures depends heavily on the perceived intent or bias of the speaker. In their experiments, they asked participants to assess, among other things, the causes of a vehicle accident, with the account of the accident being given by different parties: a neutral bystander vs. the driver of the vehicle. Their results show that the bystander’s information is more trusted, but also that *incorrect* presuppositions are recalled as ‘facts’ by the participants if the account was provided by the neutral source, but not the biased source; *even if they observed the correct facts to begin with*. Dodd and Bradshaw argue that this is because the participants filtered the information relative to their perceived intention of the person providing the account.

The Dilution Effect. Tetlock and Boettger [126] investigated the effect of implicature with respect to the information presented, particularly its relevance, showing that when presented with additional, irrelevant information, people’s implicatures are *diluted*. They performed a series of controlled experiments in which participants were presented with information about an individual David, and were asked to make predictions about David’s future; for example, what his grade point average (GPA) would be. There were two control groups and two test groups. In the control groups, people were told David spent either 3 or 31 hours studying each week (which we will call groups C3 and C31), while in the *diluted* group test groups, subjects were also provided with additional irrelevant information about David (groups T3 and T31). The results showed that those in the diluted T3 group predicted a *higher* GPA than those in the undiluted C3 group, while those in the diluted T31 group predicted a *lower* GPA than those in the undiluted C31 group. Tetlock and Boettger argued that this is because participants assumed the irrelevant information may have indeed been relevant, but its lack of support for prediction led to less extreme predictions. This study and studies on which it built demonstrate the importance of relevance in explanation.

In a further study, Tetlock et al. [127] explicitly controlled for conversational maxims, by informing one set of participants that the information displayed to them was chosen at random from the history of the individual. Their results showed that the dilution effect disappeared when conversational maxims were deactivated, providing further evidence

for the dilution effect.

Together, these bodies of work and those on which they build demonstrate that Grice's maxims are indeed important in explanation for several reasons; notably that they are a good model for how people expect conversation to happen. Further, while it is clear that providing more information than necessary not only would increase the cognitive load of the explainee, but that it dilutes the effects of the information that is truly important.

5.1.3. Linguistic structure

Malle et al. [84] argue that the linguistic structure of explanations plays an important role in interpersonal explanation. They hypothesise that some linguistic devices are used not to change the reason, but to indicate *perspective* and to *manage impressions*. They asked experimental participants to select three negative and three positive intentional actions that they did recently that were outside of their normal routine. They then asked participants to explain why they did this, and coded the answers. Their results showed several interesting findings.

First, explanations for reasons can be provided in two different ways: *marked* or *unmarked*. An unmarked reason is a direct reason, while a marked reason has a *mental state marker* attached. For example, to answer the question “*Why did she go back into the house*”, the explanations “*The key is still in the house*” and “*She thinks the key is still in the house*” both give the same reason, but with different constructs that are used to give different impressions: the second explanation gives an impression that the explainee may not be in agreement with the actor.

Second, people use *belief* markers and *desire* markers; for example, “*She thinks the key is in the house*” and “*She wants the key to be in her pocket*” respectively. In general, dropping *first-person markings*, that is, a speaker dropping “*I/we believe*”, is common in conversation and the listeners automatically infer that this is a belief of the speaker. For example, “*The key is in the house*” indicates a belief on the behalf of the speaker and inferred to mean “*I believe the key is in the house*” [84]⁷.

However, for *third-person perspective*, this is not the case. The unmarked version of explanations, especially belief markers, generally imply some sort of agreement from the explainer: “*She went back in because the key is in the house*” invites the explainee to infer that the actor and the explainer *share* the belief that the key is in the house. Whereas, “*She went back in because she believes the key is in the house*” is ambiguous — it does not (necessarily) indicate the belief of the speaker. The reason: “*She went back in because she mistakenly believes the key is in the house*” offers no ambiguity of the speaker's belief.

Malle [80, p. 169, Table 6.3] argues that different markers sit on a scale between being *distancing* to being *embracing*. For example, “*she mistakenly believes*” is more distancing than “*she jumped to the conclusion*”, while “*she realises*” is embracing. Such constructs aim not to provide different reasons, but merely allow the speaker to form impressions about themselves and the actor.

5.1.4. Argumentation and Explanation

Antaki and Leudar [1] extend Hilton's model of explanation as conversation [48] from

⁷Malle [80, Chapter 4] also briefly discusses *valuings* as markers, such as “*She likes*”, but notes that these are rarely dropped in reasons.

dialogues to *arguments*. Their research shows that a majority of statements made in explanations are actually argumentative claim-backings; that is, justifying that a particular cause indeed did hold (or was thought to have held) when a statement is made. Thus, explanations are used to both report causes, but also to *back claims*, which is an argument rather than just a question-answer model. They extend the conversational model to a wider class of contrast cases. As well as explaining causes, one must be prepared to defend a particular claim made in a causal explanation. Thus, explanations extend not just to the state of affairs external to the dialogue, but also to the internal attributes of the dialogue itself.

An example on the distinction between explanation and argument provided by Antaki and Leudar [1, p. 186] is “*The water is hot because the central heating is on*”. The distinction lies on whether the *speaker* believes that the *hearer* believes that the water is hot or not. If it is believed that the speaker believes that the water is hot, then the central heating being on offers an explanation: it contrasts with a case in which the water is not hot. If the speaker believes that the hearer does not believe the water is hot, then this is an argument that the water should indeed be hot; particularly if the speaker believes that the hearer believes that the central heating is on. The speaker is thus trying to persuade the hearer that the water is hot. However, the distinction is not always so clear because explanations can have argumentative functions.

5.2. Dialogue

If we accept model of explanation as conversation, then we may ask whether there are particular dialogue structures for explanation. There has been a collection of such articles ranging from dialogues for pragmatic explanation [132] to definitions based on transfer of understanding [134]. However, the most relevant for the problem of explanation in AI is a body of work lead largely by Walton.

Walton [135] proposed a dialectical theory of explanation, putting forward similar ideas to that of Antaki and Leudar [1] in that some parts of an explanatory dialogue require the explainer to provide backing arguments to claims. In particular, he argues that such an approach is more suited to ‘everyday’ or interpersonal explanation than models based on scientific explanation. He further argues that such models should be combined with ideas of *explanation as understanding*, meaning that social explanation is about transferring knowledge from explainer to explainee. He proposes a series of conditions on the dialogue and its interactions as to when and how an explainer should transfer knowledge to an explainee.

In a follow-on paper, Walton [137] proposes a formal dialogue model called *CE*, based on an earlier persuasion dialogue [139], which defines the conditions on how a dialogue commences, rules for governing the locutions in the dialogue, rules for governing the structure or sequence of the dialogue, success rules and termination rules.

Extending on this work further [137], Walton [138] describes an improved formal dialogue system for explanation, including a set of speech act rules for practical explanation, consisting of an opening stage, exploration stage, and closing stage. In particular, this paper focuses on the closing stage to answer the question: how do we know that an explanation has ‘finished’? Scriven [121] argues that to test someone’s understanding of a topic, merely asking them to recall facts that have been told to them is insufficient — we should also be able to answer new questions that demonstrate generalisation of and inference from what has been learnt: an *examination*.

To overcome this, Walton proposes the use of *examination dialogues* [136] as a method for the explainer to determine whether the explainee has correctly understood the explanation — that is, the explainer has a real understanding, not merely a perceived (or claimed) understanding. Walton proposes several rules for the closing stage of the examination dialogue, including a rule for terminating due to ‘practical reasons’, which aim to solve the problem of the *failure cycle*, in which repeated explanations are requested, and thus the dialogue does not terminate.

Arioua and Croitoru [2] present a formalism of Walton’s work on explanation dialogue [138], grounded in a well-known argumentation framework [108]. In addition, they provide formalisms of *commitment stores* and *understanding stores* for maintaining what each party in the dialogue is committed to, and what they already understand. This is necessary to prevent circular arguments. They further define how to shift between different dialogues in order to enable nested explanations, in which an explanation produces a new why-question, but also to shift from an explanation to an argumentation dialogue, which supports nested argument due to a challenge from an explainee, as noted by Antaki and Leudar [1]. The rules define when this dialectical shift can happen, when it can return to the explanation, and what the transfer of states is between these; that is, how the explanation state is updated after a nested argument dialogue.

5.3. Social Explanation and Artificial Intelligence

This section presents some ideas on how research from social explanation affects researchers and practitioners in AI.

5.3.1. Conversational Model

The conversational model of explanation according to Hilton [48], and its subsequent extension by Antaki and Leudar [1] to consider argumentation, are appealing and useful models for explanation in AI. In particular, they are appealing because of its generality — it can be used to explain human or agent actions, emotions, physical events, algorithmic decisions, etc. It abstracts away from the cognitive processes of causal attribution and explanation selection, and therefore does not commit to any particular model of decision making, of how causes are determined, how explanations are selected, or even any particular mode of interaction.

One may argue that in digital systems, many explanations would be better done in a visual manner, rather than a conversational manner. However, the models of Hilton [48], Antaki and Leudar [1], and Walton [138] are all quite independent of language. They define interactions based on questions and answers, but these need not be verbal. Questions could be asked by interacting with a visual object, and answers could similarly be provided in a visual way. While Grice’s maxims are about conversation, they apply just as well to other modes of interaction. For instance, a good visual explanation would display only quality explanations that are relevant and relate to the question — these are exactly Grice’s maxims.

I argue that, if we are to design and implement agents that can truly explain themselves, in many scenarios, the explanation will have to be interactive and adhere to maxims of communication, irrelevant of the media used.

5.3.2. *Model of Other*

In Section 2.6.4, I argue that an explanation-friendly *model of self* is required to provide meaningful explanations. However, for social explanation, a *model of other* is also required. Clearly, as part of a dialog, an explanatory agent should at least keep track of what has already been explained, which is a simple model of other. However, if an intelligent agent is operating with a human explainee in a particular environment, it could may have access to more complete models of other, such as the other’s capabilities and their current beliefs or knowledge; and even the explainee’s model of the explanatory agent itself. If it has such a model, the explanatory agent can exploit this by tailoring the explanation to the human observer. For example, by not explaining information that it believes the human already knows, by estimating what the human will infer from existing or new information, and even anticipating what why-questions may arise. Halpern and Pearl [36] already considers a simplified idea of this in their model of explanation, but other work on epistemic reasoning and planning [24, 101] and planning for interactive dialogue [107] can clearly play a part here.

5.3.3. *Implicature*

It is clear that in some settings, implicature can play an important role. Reasoning about implications of what the explainee says could support more succinct explanations, but just as importantly, those designing explanatory agents must also keep in mind what people could infer from the literal explanations — both correctly and incorrectly.

Further to this, as noted by Dodd and Bradshaw [20], people interpret explanations relative to the intent of the explainer. This is important for explainable AI because one of the main goals of explanation is to establish trust of people, and as such, explainees will be aware of this goal. It is clear that we should quite often assume from the outset that trust levels are low. If explainees are sceptical of the decisions made by a system, it is not difficult to imagine that they will also be sceptical of explanations provided, and could interpret explanations as biased.

5.3.4. *Dialogue*

Walton’s explanation dialogues [135, 137, 138], which build on well-accepted models from argumentation, are closer to the notion of computational models than that of Hilton [48] or Antaki and Leudar [1]. While Walton also abstracts away from the cognitive processes of causal attribution and explanation selection, his dialogues are more idealised ways of how explanation can occur, and thus make certain assumptions that may be reasonable for a model, but of course, do not account for all possible interactions. However, this is appealing from an explainable AI perspective because it is clear that the interactions between an explanatory agent and an explainee will need to be scoped to be computationally tractable. Walton’s models provide a nice step towards implementing Hilton’s conversational model.

In particular, Arioua and Croitoru’s *formal* model for explanation [2] not only brings us one step closer to a computational model, but also nicely brings together the models of Hilton [48] and Antaki and Leudar [1] for allowing arguments over claims in explanations. Such formal models of explanation could work together with concepts such as *conversation policies* [32] to implement explanations.

5.3.5. Dilution

Finally, it is important to focus on dilution. As noted in the introduction of this paper, much of the work in explainable AI is focused on causal attributions. The work outlined in Section 4 shows that this is only part of the problem. While presenting a casual chain may allow an explainee to fill in the gaps of their own knowledge, there is still a likely risk that the less relevant parts of the chain will dilute those parts that are crucial to the particular question asked by the explainee. Thus, this again emphasises the importance of explanation selection and relevance.

6. Conclusions

In this paper, I have argued that explainable AI can benefit from existing models of how people define, generate, select, present, and evaluate explanations. I have reviewed what I believe are some of the most relevant and important findings from social science research on human explanation, and have provide some insight into how this work can be used in explainable AI.

In particular, we should take the three major findings noted in the introduction into account in our explainable AI models: (1) why-questions are contrastive; (2) explanations are selected; and (3) explanations are social. I acknowledge that incorporating these ideas will not always be feasible for all applications, but in many cases, they have the potential to improve explanatory agents. I hope and expect that readers will also find other useful ideas from this survey.

The primary goal of this paper is to provide a portal to work on explainable AI that enables researchers and practitioners in this area to infuse such work into their own models. However, my intention is that the paper should be used merely as a starting point for people to explore particular areas of this literature. It is clear that adopting this work into explainable AI is not a straightforward step. From a social science viewpoint, these models will need to be refined and extended to provide good explanatory agents, which requires researchers in explainable AI to work closely with researchers from psychology and cognitive science. Already, projects of this type are underway, with impressive results; for example, see [64, 62, 117].

Explainable AI, Social Science, and Interaction Design

It is important to note that building explainable AI is not a computer science problem, but a multi-disciplinary problem. In particular, there are two other disciplines of importance: social science and interaction design.

This article highlights the philosophy, social and cognitive psychology, and cognitive science are important areas, but I contend that we should work collaboratively with researchers from these areas to both build and validate explainable AI models.

However, a link that is not touched on in this paper, is the field of *interaction design*. Much of the research into human explanation is about verbal conversation. In AI, explanation will often be well-supported by other media, such as visual aids, offering a more powerful interaction than via text alone. But it is important not to fall into the trap of thinking that we can independently build explanatory agents and then just hand it over to an interaction designer to figure out how to present the results to the observer. For example, in Section 2.6.4, I discuss the ‘model of self’ that is required to provide explanations. In many applications, this model of self will be tightly coupled to the particular

way that the observer interacts with the explanation, and thus, the interaction design must be considered in conjunction with the design of the model of self. Similarly, the ‘model of other’ discussed in Section 5.3.2 is related to the interaction in general. We may infer what the human observer already knows by using gaze tracking technology, for example. As such, the design of the model of other will be tightly integrated with the interaction design in many applications.

I hope that readers agree with this position, and consider adopting a more multi-disciplinary approach to explainable AI to provide agents that are truly able to provide meaningful explanations to people.

Acknowledgements

The author would like to thank Denis Hilton for his review on an earlier draft of this paper, pointers to several pieces of related work, and for his many insightful discussions on the link between explanation in social sciences and artificial intelligence. The author would also like to thank Natasha Goss for her critical review of an earlier draft of this paper, and Darryn Reid for his discussions on the link between self, trust, and explanation, and for the term ‘Model of Self’.

This work was undertaken while the author was on sabbatical at the Université de Toulouse Capitole, and was partially funded by Australian Research Council DP160104083 *Catering for individuals’ emotions in technology development* and Defence Science and Technology CERA grant 001 *Causal Explanation in Trusted Autonomous Systems*.

References

- [1] C. Antaki, I. Leudar, Explaining in conversation: Towards an argument model, *European Journal of Social Psychology* 22 (2) (1992) 181–194.
- [2] A. Arioua, M. Croitoru, Formalizing explanatory dialogues, in: *International Conference on Scalable Uncertainty Management*, Springer, 282–297, 2015.
- [3] J. L. Aronson, On the grammar of ‘cause’, *Synthese* 22 (3) (1971) 414–430.
- [4] P. Besnard, A. Hunter, *Elements of argumentation*, vol. 47, MIT press Cambridge, 2008.
- [5] A. Boonzaier, J. McClure, R. M. Sutton, Distinguishing the effects of beliefs and preconditions: The folk psychology of goals and actions, *European Journal of Social Psychology* 35 (6) (2005) 725–740.
- [6] R. I. Brafman, C. Domshlak, From One to Many: Planning for Loosely Coupled Multi-Agent Systems., in: *International Conference on Automated Planning and Scheduling*, 28–35, 2008.
- [7] S. Bromberger, Why-questions, in: R. G. Colodny (Ed.), *Mind and Cosmos: Essays in Contemporary Science and Philosophy*, Pittsburgh University Press, Pittsburgh, 68–111, 1966.
- [8] B. Buchanan, E. Shortliffe, *Rule-based expert systems: the MYCIN experiments of the Stanford Heuristic Programming Project*, Addison-Wesley, 1984.
- [9] A. Burguet, D. Hilton, Effets de contexte sur l’explication causale, in: M. B. et A. Trognon (Ed.), *Psychologie Sociale et Communication*, Paris: Dunod, 219–228, 2004.
- [10] R. M. Byrne, The Construction of Explanations, in: *AI and Cognitive Science’90*, Springer, 337–351, 1991.
- [11] K. Chan, T.-W. Lee, P. A. Sample, M. H. Goldbaum, R. N. Weinreb, T. J. Sejnowski, Comparison of machine learning and traditional classifiers in glaucoma diagnosis, *IEEE Transactions on Biomedical Engineering* 49 (9) (2002) 963–974.
- [12] B. Chandrasekaran, M. C. Tanner, J. R. Josephson, Explaining control strategies in problem solving, *IEEE Expert* 4 (1) (1989) 9–15.
- [13] J. Y. Chen, K. Procci, M. Boyce, J. Wright, A. Garcia, M. Barnes, Situation awareness-based agent transparency, Tech. Rep. ARL-TR-6905, U.S. Army Research Laboratory, 2014.
- [14] Y. Chevalayre, U. Endriss, J. Lang, N. Maudet, A short introduction to computational social choice, in: *International Conference on Current Trends in Theory and Practice of Computer Science*, Springer, 51–69, 2007.

- [15] S. Chin-Parker, A. Bradner, Background shifts affect explanatory style: how a pragmatic theory of explanation accounts for background effects in the generation of explanations, *Cognitive Processing* 11 (3) (2010) 227–249.
- [16] H. Chockler, J. Y. Halpern, Responsibility and blame: A structural-model approach, *Journal of Artificial Intelligence Research* 22 (2004) 93–115.
- [17] A. Cooper, The inmates are running the asylum: Why high-tech products drive us crazy and how to restore the sanity, Sams Indianapolis, IN, USA, 2004.
- [18] DARPA, Explainable Artificial Intelligence (XAI) Program, <http://www.darpa.mil/program/explainable-artificial-intelligence>, full solicitation at <http://www.darpa.mil/attachments/DARPA-BAA-16-53.pdf>, 2016.
- [19] D. C. Dennett, *The intentional stance*, MIT press, 1989.
- [20] D. H. Dodd, J. M. Bradshaw, Leading questions and memory: Pragmatic constraints, *Journal of Memory and Language* 19 (6) (1980) 695.
- [21] P. Dowe, Wesley Salmon’s process theory of causality and the conserved quantity theory, *Philosophy of Science* 59 (2) (1992) 195–216.
- [22] T. Eiter, T. Lukasiewicz, Complexity results for structure-based causality, *Artificial Intelligence* 142 (1) (2002) 53–89.
- [23] T. Eiter, T. Lukasiewicz, Causes and explanations in the structural-model approach: Tractable cases, *Artificial Intelligence* 170 (6-7) (2006) 542–580.
- [24] R. Fagin, J. Halpern, Y. Moses, M. Vardi, *Reasoning about knowledge*, vol. 4, MIT press Cambridge, 1995.
- [25] D. Fair, Causation and the Flow of Energy, *Erkenntnis* 14 (3) (1979) 219–250.
- [26] J. Fox, D. Glasspool, D. Grecu, S. Modgil, M. South, V. Patkar, Argumentation-based inference and decision making—A medical perspective, *IEEE intelligent systems* 22 (6).
- [27] T. Gerstenberg, D. A. Lagnado, Spreading the blame: The allocation of responsibility amongst multiple agents, *Cognition* 115 (1) (2010) 166–171.
- [28] M. Ghallab, D. Nau, P. Traverso, *Automated Planning: theory and practice*, Elsevier, 2004.
- [29] C. Ginet, In defense of a non-causal account of reasons explanations, *The Journal of Ethics* 12 (3-4) (2008) 229–237.
- [30] L. Giordano, C. Schwind, Conditional logic of actions and causation, *Artificial Intelligence* 157 (1-2) (2004) 239–279.
- [31] V. Girotto, P. Legrenzi, A. Rizzo, Event controllability in counterfactual thinking, *Acta Psychologica* 78 (1) (1991) 111–133.
- [32] M. Greaves, H. Holmback, J. Bradshaw, What is a conversation policy?, in: *Issues in Agent Communication*, Springer, 118–131, 2000.
- [33] H. P. Grice, Logic and conversation, in: *Syntax and semantics 3: Speech arts*, New York: Academic Press, 41–58, 1975.
- [34] J. Y. Halpern, Axiomatizing causal reasoning, *Journal of Artificial Intelligence Research* 12 (2000) 317–337.
- [35] J. Y. Halpern, J. Pearl, Causes and explanations: A structural-model approach. Part I: Causes, *The British Journal for the Philosophy of Science* 56 (4) (2005) 843–887.
- [36] J. Y. Halpern, J. Pearl, Causes and explanations: A structural-model approach. Part II: Explanations, *The British Journal for the Philosophy of Science* 56 (4) (2005) 889–911.
- [37] R. J. Hankinson, *Cause and explanation in ancient Greek thought*, Oxford University Press, 2001.
- [38] N. R. Hanson, *Patterns of discovery: An inquiry into the conceptual foundations of science*, CUP Archive, 1965.
- [39] H. L. A. Hart, T. Honoré, *Causation in the Law*, OUP Oxford, 1985.
- [40] B. Hayes, J. A. Shah, Improving Robot Controller Transparency Through Autonomous Policy Explanation, in: *Proceedings of the 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI 2017)*, 2017.
- [41] F. Heider, *The psychology of interpersonal relations*, New York: Wiley, 1958.
- [42] F. Heider, M. Simmel, An experimental study of apparent behavior, *The American Journal of Psychology* 57 (2) (1944) 243–259.
- [43] C. G. Hempel, P. Oppenheim, Studies in the Logic of Explanation, *Philosophy of Science* 15 (2) (1948) 135–175.
- [44] G. Hesslow, The problem of causal selection, *Contemporary science and natural explanation: Commonsense conceptions of causality* (1988) 11–32.
- [45] D. Hilton, Social Attribution and Explanation, in: *Oxford Handbook of Causal Reasoning*, Oxford University Press, 645–676, 2017.

- [46] D. Hilton, C. Schmeltzer, W. C., D. Mandel, Do causal attributions depend on counterfactuals? Distinguishing *but-for* from *if-only* tests, in: Counterfactual thinking: Functions, causes, emotions, 2016.
- [47] D. J. Hilton, Logic and causal attribution, in: Contemporary science and natural explanation: Commonsense conceptions of causality, New York University Press, 33–65, 1988.
- [48] D. J. Hilton, Conversational processes and causal explanation, Psychological Bulletin 107 (1) (1990) 65–81.
- [49] D. J. Hilton, Mental models and causal explanation: Judgements of probable cause and explanatory relevance, Thinking & Reasoning 2 (4) (1996) 273–308.
- [50] D. J. Hilton, J. McClure, B. Slugoski, Counterfactuals, conditionals and causality: A social psychological perspective, in: D. R. Mande, D. J. Hilton, P. Catellani (Eds.), The psychology of counterfactual thinking, London: Routledge, 44–60, 2005.
- [51] D. J. Hilton, J. McClure, R. M. Sutton, Selecting explanations from causal chains: Do statistical principles explain preferences for voluntary causes?, European Journal of Social Psychology 40 (3) (2010) 383–400.
- [52] D. J. Hilton, J. L. McClure, R. Slugoski, Ben, The Course of Events: Counterfactuals, Causal Sequences and Explanation, in: D. R. Mandel, D. J. Hilton, P. Catellani (Eds.), The Psychology of Counterfactual Thinking, Routledge, 2005.
- [53] D. J. Hilton, B. R. Slugoski, Knowledge-based causal attribution: The abnormal conditions focus model, Psychological review 93 (1) (1986) 75.
- [54] D. Hume, T. L. Beauchamp, An enquiry concerning human understanding: A critical edition, vol. 3, Oxford University Press, 2000.
- [55] J. M. Jaspars, D. J. Hilton, Mental models of causal reasoning, in: The social psychology of knowledge, Cambridge University Press, 335–358, 1988.
- [56] D. Kahneman, A. Tversky, The simulation heuristic, in: P. S. D. Kahneman, A. Tversky (Eds.), Judgment under Uncertainty: Heuristics and Biases, New York: Cambridge University Press, 1982.
- [57] Y. Kashima, A. McKintyre, P. Clifford, The category of the mind: Folk psychology of belief, desire, and intention, Asian Journal of Social Psychology 1 (3) (1998) 289–313.
- [58] A. Kass, D. Leake, Types of Explanations, Tech. Rep. ADA183253, DTIC Document, 1987.
- [59] H. H. Kelley, Attribution theory in social psychology, in: Nebraska symposium on motivation, University of Nebraska Press, 192–238, 1967.
- [60] H. H. Kelley, Causal schemata and the attribution process, General Learning Press, Morristown, NJ, 1972.
- [61] J. Knobe, Intentional action and side effects in ordinary language, Analysis 63 (279) (2003) 190–194.
- [62] T. Kulesza, M. Burnett, W.-K. Wong, S. Stumpf, Principles of explanatory debugging to personalize interactive machine learning, in: Proceedings of the 20th International Conference on Intelligent User Interfaces, ACM, 126–137, 2015.
- [63] T. Kulesza, S. Stumpf, M. Burnett, S. Yang, I. Kwan, W.-K. Wong, Too much, too little, or just right? Ways explanations impact end users’ mental models, in: Visual Languages and Human-Centric Computing (VL/HCC), 2013 IEEE Symposium on, IEEE, 3–10, 2013.
- [64] T. Kulesza, S. Stumpf, W.-K. Wong, M. M. Burnett, S. Perona, A. Ko, I. Oberst, Why-oriented end-user debugging of naive Bayes text classification, ACM Transactions on Interactive Intelligent Systems (TiiS) 1 (1) (2011) 2.
- [65] D. A. Lagnado, S. Channon, Judgments of cause and blame: The effects of intentionality and foreseeability, Cognition 108 (3) (2008) 754–770.
- [66] P. Langley, B. Meadows, M. Sridharan, D. Choi, Explainable Agency for Intelligent Autonomous Systems, in: Proceedings of the Twenty-Ninth Annual Conference on Innovative Applications of Artificial Intelligence, AAAI Press, 2017.
- [67] J. Leddo, R. P. Abelson, P. H. Gross, Conjunctive explanations: When two reasons are better than one, Journal of Personality and Social Psychology 47 (5) (1984) 933.
- [68] D. Lewis, Causation, The Journal of Philosophy 70 (17) (1974) 556–567.
- [69] D. Lewis, Causal explanation, Philosophical Papers 2 (1986) 214–240.
- [70] M. P. Linegang, H. A. Stoner, M. J. Patterson, B. D. Seppelt, J. D. Hoffman, Z. B. Crittendon, J. D. Lee, Human-automation collaboration in dynamic mission planning: A challenge requiring an ecological approach, Proceedings of the Human Factors and Ergonomics Society Annual Meeting 50 (23) (2006) 2482–2486.
- [71] P. Lipton, Contrastive explanation, Royal Institute of Philosophy Supplement 27 (1990) 247–266.

- [72] Z. C. Lipton, The mythos of model interpretability, arXiv preprint arXiv:1606.03490 .
- [73] T. Lombrozo, The structure and function of explanations, *Trends in Cognitive Sciences* 10 (10) (2006) 464–470.
- [74] T. Lombrozo, Simplicity and probability in causal explanation, *Cognitive psychology* 55 (3) (2007) 232–257.
- [75] T. Lombrozo, Explanation and categorization: How “why?” informs “what?”, *Cognition* 110 (2) (2009) 248–253.
- [76] T. Lombrozo, Causal–explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions, *Cognitive Psychology* 61 (4) (2010) 303–332.
- [77] T. Lombrozo, Explanation and abductive inference, *Oxford handbook of thinking and reasoning* (2012) 260–276.
- [78] J. L. Mackie, *The cement of the universe*, Oxford, 1980.
- [79] B. F. Malle, How people explain behavior: A new theoretical framework, *Personality and Social Psychology Review* 3 (1) (1999) 23–48.
- [80] B. F. Malle, *How the mind explains behavior: Folk explanations, meaning, and social interaction*, MIT Press, 2004.
- [81] B. F. Malle, Attribution theories: How people make sense of behavior, *Theories in Social Psychology* (2011) 72–95.
- [82] B. F. Malle, Time to Give Up the Dogmas of Attribution: An Alternative Theory of Behavior Explanation, *Advances in Experimental Social Psychology* 44 (1) (2011) 297–311.
- [83] B. F. Malle, J. Knobe, The folk concept of intentionality, *Journal of Experimental Social Psychology* 33 (2) (1997) 101–121.
- [84] B. F. Malle, J. Knobe, M. J. O’laughlin, G. E. Pearce, S. E. Nelson, Conceptual structure and social functions of behavior explanations: Beyond person–situation attributions, *Journal of Personality and Social Psychology* 79 (3) (2000) 309.
- [85] B. F. Malle, J. M. Knobe, S. E. Nelson, Actor-observer asymmetries in explanations of behavior: new answers to an old question, *Journal of Personality and Social Psychology* 93 (4) (2007) 491.
- [86] B. F. Malle, G. E. Pearce, Attention to behavioral events during interaction: Two actor-observer gaps and three attempts to close them, *Journal of Personality and Social Psychology* 81 (2) (2001) 278–294.
- [87] D. Marr, *Vision: A computational investigation into the human representation and processing of visual information*, Inc., New York, NY, 1982.
- [88] D. Marr, T. Poggio, From understanding computation to understanding neural circuitry, *AI Memos AIM-357*, MIT, 1976.
- [89] R. McCloy, R. M. Byrne, Counterfactual thinking about controllable events, *Memory & Cognition* 28 (6) (2000) 1071–1078.
- [90] J. McClure, Goal-based explanations of actions and outcomes, *European Review of Social Psychology* 12 (1) (2002) 201–235.
- [91] J. McClure, D. Hilton, For you can’t always get what you want: When preconditions are better explanations than goals, *British Journal of Social Psychology* 36 (2) (1997) 223–240.
- [92] J. McClure, D. Hilton, J. Cowan, L. Ishida, M. Wilson, When rich or poor people buy expensive objects: Is the question how or why?, *Journal of Language and Social Psychology* 20 (2001) 229–257.
- [93] J. McClure, D. J. Hilton, Are goals or preconditions better explanations? It depends on the question, *European Journal of Social Psychology* 28 (6) (1998) 897–911.
- [94] J. L. McClure, R. M. Sutton, D. J. Hilton, The Role of Goal-Based Explanations, in: *Social judgments: Implicit and explicit processes*, vol. 5, Cambridge University Press, 306, 2003.
- [95] A. L. McGill, J. G. Klein, Contrastive and counterfactual reasoning in causal judgment, *Journal of Personality and Social Psychology* 64 (6) (1993) 897.
- [96] P. Menzies, H. Price, Causation as a secondary quality, *The British Journal for the Philosophy of Science* 44 (2) (1993) 187–203.
- [97] J. E. Mercado, M. A. Rupp, J. Y. Chen, M. J. Barnes, D. Barber, K. Procci, Intelligent agent transparency in human–agent teaming for Multi-UxV management, *Human Factors* 58 (3) (2016) 401–415.
- [98] J. S. Mill, *A system of logic: The collected works of John Stuart Mill*, vol. III, 1973.
- [99] D. T. Miller, S. Gunasegaram, Temporal order and the perceived mutability of events: Implications for blame assignment, *Journal of personality and social psychology* 59 (6) (1990) 1111.
- [100] T. Miller, P. Howe, L. Sonenberg, Explainable AI: Beware of Inmates Running the Asylum, in: *IJCAI 2017 Workshop on Explainable Artificial Intelligence (XAI)*, 1–5, URL <http://people.>

- unimelb.edu.au/tmiller/pubs/explanation-inmates.pdf, (In Press), 2017.
- [101] C. Muise, V. Belle, P. Felli, S. McIlraith, T. Miller, A. R. Pearce, L. Sonenberg, Planning Over Multi-Agent Epistemic States: A Classical Planning Approach, in: B. Bonet, S. Koenig (Eds.), Proceedings of AAAI 2015, 1–8, 2015.
 - [102] G. Nott, ‘Explainable Artificial Intelligence’: Cracking open the black box of AI, Computer World <https://www.computerworld.com.au/article/617359/>.
 - [103] M. J. O’laughlin, B. F. Malle, How people explain actions performed by groups and individuals, *Journal of Personality and Social Psychology* 82 (1) (2002) 33.
 - [104] J. Overton, Scientific Explanation and Computation, in: D. B. L. Thomas Roth-Berghofer, Nava Tintarev (Ed.), Proceedings of the 6th International Explanation-Aware Computing (Ex-aCt) workshop, 41–50, 2011.
 - [105] J. A. Overton, Explanation in Science, Ph.D. thesis, The University of Western Ontario, 2012.
 - [106] J. A. Overton, “Explain” in scientific discourse, *Synthese* 190 (8) (2013) 1383–1405.
 - [107] R. Petrick, M. E. Foster, Using General-Purpose Planning for Action Selection in Human-Robot Interaction, in: AAAI 2016 Fall Symposium on Artificial Intelligence for Human-Robot Interaction, 2016.
 - [108] H. Prakken, Formal systems for persuasion dialogue, *The Knowledge Engineering Review* 21 (02) (2006) 163–188.
 - [109] S. Prasada, The scope of formal explanation, *Psychonomic Bulletin & Review* (2017) 1–10.
 - [110] S. Prasada, E. M. Dillingham, Principled and statistical connections in common sense conception, *Cognition* 99 (1) (2006) 73–112.
 - [111] J. Preston, N. Epley, Explanations versus applications: The explanatory power of valuable beliefs, *Psychological Science* 16 (10) (2005) 826–832.
 - [112] M. Ranney, P. Thagard, Explanatory coherence and belief revision in naive physics, in: Proceedings of the Tenth Annual Conference of the Cognitive Science Society, 426–432, 1988.
 - [113] A. S. Rao, M. P. Georgeff, BDI agents: From theory to practice., in: ICMAS, vol. 95, 312–319, 1995.
 - [114] S. J. Read, A. Marcus-Newhall, Explanatory coherence in social explanations: A parallel distributed processing account, *Journal of Personality and Social Psychology* 65 (3) (1993) 429.
 - [115] B. Rehder, A causal-model theory of conceptual representation and categorization, *Journal of Experimental Psychology: Learning, Memory, and Cognition* 29 (6) (2003) 1141.
 - [116] B. Rehder, When similarity and causality compete in category-based property generalization, *Memory & Cognition* 34 (1) (2006) 3–16.
 - [117] M. T. Ribeiro, S. Singh, C. Guestrin, Why Should I Trust You?: Explaining the Predictions of Any Classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 1135–1144, 2016.
 - [118] W. C. Salmon, Four decades of scientific explanation, University of Pittsburgh press, 2006.
 - [119] J. Samland, M. Josephs, M. R. Waldmann, H. Rakoczy, The role of prescriptive norms and knowledge in children’s and adults’ causal selection, *Journal of Experimental Psychology: General* 145 (2) (2016) 125.
 - [120] J. Samland, M. R. Waldmann, Do Social Norms Influence Causal Inferences?, in: P. Bello, M. Guarini, M. McShane, B. Scassellati (Eds.), Proceedings of the 36th Annual Conference of the Cognitive Science Society, Cognitive Science Society, 1359–1364, 2014.
 - [121] M. Scriven, The concept of comprehension: From semantics to software, in: J. B. Carroll, R. O. Freedle (Eds.), *Language comprehension and the acquisition of knowledge*, Washington: W. H. Winston & Sons, 31–39, 1972.
 - [122] B. R. Slugoski, M. Lalljee, R. Lamb, G. P. Ginsburg, Attribution in conversational context: Effect of mutual knowledge on explanation-giving, *European Journal of Social Psychology* 23 (3) (1993) 219–238.
 - [123] K. Stubbs, P. Hinds, D. Wettergreen, Autonomy and common ground in human-robot interaction: A field study, *IEEE Intelligent Systems* 22 (2) (2007) 42–50.
 - [124] J. Susskind, K. Maurer, V. Thakkar, D. L. Hamilton, J. W. Sherman, Perceiving individuals and groups: expectancies, dispositional inferences, and causal attributions, *Journal of Personality and Social Psychology* 76 (2) (1999) 181.
 - [125] W. R. Swartout, J. D. Moore, Explanation in second generation expert systems, in: *Second Generation Expert Systems*, Springer, 543–585, 1993.
 - [126] P. E. Tetlock, R. Boettger, Accountability: a social magnifier of the dilution effect, *Journal of Personality and Social Psychology* 57 (3) (1989) 388.
 - [127] P. E. Tetlock, J. S. Learner, R. Boettger, The dilution effect: judgemental bias, conversational

- convention, or a bit of both?, *European Journal of Social Psychology* 26 (1996) 915–934.
- [128] P. Thagard, Explanatory coherence, *Behavioral and Brain Sciences* 12 (03) (1989) 435–467.
 - [129] T. Trabasso, J. Bartolone, Story understanding and counterfactual reasoning, *Journal of Experimental Psychology: Learning, Memory, and Cognition* 29 (5) (2003) 904.
 - [130] A. Tversky, D. Kahneman, Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment, *Psychological Review* 90 (4) (1983) 293.
 - [131] K. Uttich, T. Lombrozo, Norms inform mental state ascriptions: A rational explanation for the side-effect effect, *Cognition* 116 (1) (2010) 87–100.
 - [132] B. C. Van Fraassen, The pragmatics of explanation, *American Philosophical Quarterly* 14 (2) (1977) 143–150.
 - [133] N. Vasilyeva, D. A. Wilkenfeld, T. Lombrozo, Goals Affect the Perceived Quality of Explanations., in: D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, P. P. Maglio (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society*, Cognitive Science Society, 2469–2474, 2015.
 - [134] G. H. Von Wright, *Explanation and understanding*, Cornell University Press, 1971.
 - [135] D. Walton, A new dialectical theory of explanation, *Philosophical Explorations* 7 (1) (2004) 71–89.
 - [136] D. Walton, Examination dialogue: An argumentation framework for critically questioning an expert opinion, *Journal of Pragmatics* 38 (5) (2006) 745–777.
 - [137] D. Walton, Dialogical Models of Explanation, in: *Proceedings of the International Explanation-Aware Computing (ExaCt) workshop*, 1–9, 2007.
 - [138] D. Walton, A dialogue system specification for explanation, *Synthese* 182 (3) (2011) 349–374.
 - [139] D. N. Walton, *Logical Dialogue — Games and Fallacies*, University Press of America, Lanham, Maryland, 1984.
 - [140] A. Wendt, On constitution and causation in international relations, *Review of International Studies* 24 (05) (1998) 101–118.
 - [141] J. J. Williams, T. Lombrozo, B. Rehder, The hazards of explanation: Overgeneralization in the face of exceptions, *Journal of Experimental Psychology: General* 142 (4) (2013) 1006.
 - [142] J. Woodward, *Making things happen: A theory of causal explanation*, Oxford University Press, 2005.