Artificial Intelligence 267 (2019) 1–38

Contents lists available at ScienceDirect

Artificial Intelligence

www.elsevier.com/locate/artint

Explanation in artificial intelligence: Insights from the social sciences

Tim Miller

School of Computing and Information Systems, University of Melbourne, Melbourne, Australia

ARTICLE INFO

Article history:
Received 22 June 2017
Received in revised form 17 May 2018
Accepted 16 July 2018
Available online 27 October 2018

Keywords:
Explanation
Explainability
Interpretability
Explainable AI
Transparency

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 looking at how humans explain to each other can serve as a useful starting point for explanation in artificial intelligence. 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 exist vast and valuable bodies of research in philosophy, psychology, and cognitive science of how people define, generate, select, evaluate, and present explanations, which argues that people employ certain cognitive biases and social expectations to the explanation process. This paper argues that the field of explainable artificial intelligence can 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.

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. [23], [168], and Buchanan and Shortliffe [14]. Sometimes abbreviated XAI (eXplainable artificial intelligence), the idea can be found in grant solicitations [32] and in the popular press [136]. This resurgence is driven by evidence that many AI applications have limited take up, or are not appropriated at all, due to ethical concerns [2] and a *lack of trust* on behalf of their users [166,101]. The running hypothesis is that by building more transparent, interpretable, or explainable systems, users will be better equipped to understand and therefore trust the intelligent agents [129,25,65].

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 decisions[1] 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 [129,65], debugging of machine learning models [89], explaining medical decision-making [45], and explaining predictions of classifiers [157].

E-mail address: tmiller@unimelb.edu.au.

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

https://doi.org/10.1016/j.artint.2018.07.007
0004-3702/

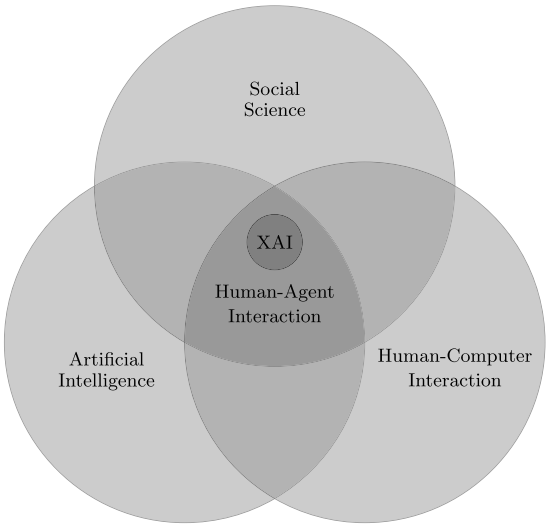

Fig. 1. Scope of explainable artificial intelligence.

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 decisions and behaviour to each other are a good way to start analysing the problem. Researchers argue that people employ certain *biases* [82] and *social expectations* [72] when they generate and evaluate explanation, and I argue that such biases and expectations can improve human interactions with explanatory AI. For example, de Graaf and Malle [34] argues that because people assign human-like traits to artificial agents, people will expect explanations using the same conceptual framework used to explain human behaviours.

Despite the recent resurgence of explanation and interpretability in AI, most of the research and practice in this area seems to use the researchers' intuitions of what constitutes a 'good' explanation. Miller et al. [132] shows in a small sample that research in explainable AI typically does not cite or build on frameworks of explanation from social science. They argue that 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 [31]) 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 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 can form the basis of explainable AI — although there is still a lot to learn from early work in explainable AI around design and implementation.

This paper aims to promote the inclusion of this existing research into the field of explanation in AI. As part of this work, over 250 publications on explanation were surveyed from social science venues. A smaller subset of these were chosen to be presented in this paper, based on their currency and relevance to the topic. The paper presents relevant theories on explanation, describes, in many cases, the experimental evidence supporting these theories, and presents ideas on how this work can be infused into explainable AI.

1.1. Scope

In this article, the term '*Explainable AI*' loosely refers to an explanatory agent revealing underlying causes to its or another agent's decision making. However, it is important to note that the solution to explainable AI is not just 'more AI'. Ultimately, it is a human–agent interaction problem. Human-agent interaction can be defined as the intersection of artificial intelligence, social science, and human–computer interaction (HCI); see Fig. 1. Explainable AI is just one problem inside human–agent interaction.

This article highlights the top circle in Fig. 1: the philosophy, social and cognitive psychology, and cognitive science views of explanation, and their relation to the other two circles: their impact on the design of both artificial intelligence and our interactions with them. With this scope of explainable AI in mind, the scope of this article is 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' (or local) 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 [166,129], 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 explainable AI*: To draw important points from relevant articles to some of the different sub-fields of explainable AI.

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 build on social science. For an excellent short survey on explanation in machine learning, see Biran and Cotton [9].

1.2. Major findings

As part of this review, I highlight four 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. In Sections 2–4, models of how people provide contrastive explanations are reviewed.
2. Explanation are *selected* (in a biased manner) — 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. However, this selection is influenced by certain cognitive biases. In Section 4, models of how people select explanations, including how this relates to contrast cases, are reviewed.
3. Probabilities probably don't matter — while truth and likelihood are important in explanation and probabilities really do matter, *referring* to probabilities or statistical relationships in explanation is not as effective as referring to causes. The most likely explanation is not always the *best* explanation for a person, and importantly, using statistical generalisations to explain why events occur is unsatisfying, unless accompanied by an underlying *causal* explanation for the generalisation itself.
4. Explanations are *social* — they are a transfer of knowledge, presented as part of a conversation[2] or interaction, and are thus presented relative to the explainer's beliefs about the explainee's beliefs. In Section 5, models of how people interact regarding explanations are reviewed.

These four points all converge around a single point: explanations are not just the presentation of associations and causes (*causal attribution*), they are *contextual*. While an event may have many causes, often the explainee cares only about a small subset (relevant to the context), 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,

[2] Note 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.

Table 1
A simple lay model for distinguishing common arthropods.

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

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

Fig. 2. Example Explanation Dialogue between a Person and an Explanation Agent.

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.

Now, consider the idealised and simple dialogue between a human user and 'ExplAgent', who is the interactive explanation agent, outlined in Fig. 2. 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.

We will refer back to this example throughout the paper and link difference 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 [99, p. 217].

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

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 [79, 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 [98], the definition due to Hume is in fact about *counterfactuals*, rather than dependence alone. Hume 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; cf. Lewis [98], Hilton [71].

This *classical counterfactual* model of causality is well understood but competing definitions exist. *Interventionist* theories of causality [191,58] 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 [128].

Transference theories [5,43,39], on the other hand, are not defined on dependence, but instead describe *physical* causation as the transference of energy between objects. In short, if E is an event representing the change of energy of an object O, then C causes E if object O is in contact with the object that causes C, and there is some *quantity* of energy transferred.

While the aim here is not a detailed survey 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 'counterfactual' is important in causality.

Gerstenberg et al. [49] tested whether people consider counterfactuals when making causal judgements in an experiment involving colliding balls. They presented experiment participants with different scenarios involving two balls colliding, with each scenario having different outcomes, such as one ball going through a gate, just missing the gate, or missing the gate by a long distance. While wearing eye-tracking equipment, participants were asked to determine what the outcome would have been (a counterfactual) had the candidate cause not occurred (the balls had not collided). Using the eye-gaze data from the tracking, they showed that their participants, even in these physical environments, would trace where the ball would have gone had the balls not collided, thus demonstrating that they used counterfactual simulation to make causal judgements.

Necessary and sufficient causes. Kelley [87] 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 [66], 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. [76] define five different types of causal chain, outlined in Table 2, and note that different causal chains are associated with different types of explanations.

People do not need to understand a complete causal chain to provide a sound explanation. This is evidently true: causes of 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 [53,98], 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 [58]. 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 [57] and a series articles on complexity analysis [40,41].

Halpern and Pearl [58] 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 of values to variables. Intuitively, a context represents a

Table 2
Types of causal chains according to Hilton et al. [76].

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 *enables* 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.

'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

"*An explanation is an assignment of causal responsibility*" — Josephson and Josephson [81].

Explanation is both a process and a product, as noted by Lombrozo [104]. 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 'filling the gaps' [27] to determine 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*).
   In social science, the process of identifying the causes of a particular phenomenon is known as *attribution*, and is seen as just *part* of the entire process of explanation.
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.

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 causes [159,191,107,59]. There are, however, definitions of non-causal explanation [52], such as explaining 'what happened' or explaining what was meant by a particular remark [187]. 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 explanation is an answer to a *why-question* [35,138,99,102].

According to Bromberger [13], 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*.

However, other types of questions can be answered by explanations. In Table 3, I propose a model for explanatory questions based on Pearl and Mackie's *Ladder of Causation* [141]. This model places explanatory questions into three classes:

Table 3
Classes of explanatory question and the reasoning required to answer.

Question	Reasoning	Description
What?	Associative	Reason about which unobserved events could have occurred given the observed events
How?	Interventionist	Simulate a change in the situation to see if the event still happens
Why?	Counterfactual	Simulating alternative causes to see whether the event still happens

(1) *what*-questions, such as "*What event happened?*"; (2) *how*-questions, such as "*How did that event happen?*"; and (3) *why*-questions, such as "*Why did that event happen?*". From the perspective of reasoning, *why*-questions are the most challenging, because they use the most sophisticated reasoning. *What*-questions ask for factual accounts, possibly using associative reasoning to determine, from the observed events, which unobserved events also happened. *How* questions are also factual, but require interventionist reasoning to determine the set of causes that, if removed, would prevent the event from happening. This may also require associative reasoning. We categorise *what if*-questions has *how*-questions, as they are just a contrast case analysing what would happen under a different situation. *Why*-questions are the most challenging, as they require counterfactual reasoning to undo events and simulate other events that are not factual. This also requires associative and interventionist reasoning.

Dennett [36] argues that "why" is ambiguous and that there are two different senses of why-question: *how come?* and *what for?*. The former asks for a *process narrative*, without an explanation of what it is *for*, while the latter asks for a *reason*, which implies some intentional thought behind the cause. Dennett gives the examples of "why are planets spherical?" and "why are ball bearings spherical?". The former asks for an explanation based on physics and chemistry, and is thus a how-come-question, because planets are not round *for* any reason. The latter asks for an explanation that gives the reason what the designer made ball bearings spherical *for*: a reason because people design them that way.

Given a why-question, Overton [138] defines an *explanation* as a pair consisting of: (1) the *explanans*: which is the answer to the question; and (2) and the *explanandum*; which is the presupposition.

2.1.4. Explanation as abductive reasoning

As a cognitive process, explanation is closely related to *abductive reasoning*. Peirce [142] was the first author to consider abduction as a distinct form of reasoning, separate from induction and deduction, but which, like induction, went from effect to cause. His work focused on the difference between accepting a hypothesis via scientific experiments (induction), and *deriving* a hypothesis to explain observed phenomenon (abduction). He defines the form of inference used in abduction as follows:

> The surprising fact, C, is observed;
> But if A were true, C would be a matter of course,
> Hence, there is reason to suspect that A is true.

Clearly, this is an inference to *explain* the fact C from the hypothesis A, which is different from deduction and induction. However, this does not account for competing hypotheses. Josephson and Josephson [81] describe this more competitive-form of abduction as:

> D is a collection of data (facts, observations, givens).
> H explains D (would, if true, explain D).
> No other hypothesis can explain D as well as H does.
> Therefore, H is probably true.

Harman [62] labels this process "*inference to the best explanation*". Thus, one can think of abductive reasoning as the following process: (1) observe some (presumably unexpected or surprising) events; (2) generate one or more hypothesis about these events; (3) judge the plausibility of the hypotheses; and (4) select the 'best' hypothesis as the explanation [78].

Research in philosophy and cognitive science has argued that abductive reasoning is closely related to explanation. In particular, in trying to understand causes of events, people use abductive inference to determine what they consider to be the "best" explanation. Harman [62] is perhaps the first to acknowledge this link, and more recently, experimental evaluations have demonstrated it [108,188,109,154]. Popper [146] is perhaps the most influential proponent of abductive reasoning in the scientific process. He argued strongly for the scientific method to be based on empirical falsifiability of hypotheses, rather than the classic inductivist view at the time.

Early philosophical work considered abduction as some magical process of intuition — something that could not be captured by formalised rules because it did not fit the standard deductive model. However, this changed when artificial intelligence researchers began investigating abductive reasoning to explain observations, such as in diagnosis (e.g. medical diagnosis, fault diagnosis) [145,156], intention/plan recognition [24], etc. The necessity to encode the process in a suitable computational form led to axiomatisations, with Pople [145] seeming to be the first to do this, and characterisations of how to implement such axiomatisations; e.g. Levesque [97]. From here, the process of abduction as a principled process gained traction, and it is now widely accepted that abduction, induction, and deduction are different modes of logical reasoning.

In this paper, abductive inference is not equated directly to explanation, because explanation also refers to the product and the social process; but abductive reasoning does fall into the category of cognitive process of explanation. In Section 4, we survey the cognitive science view of abductive reasoning, in particular, cognitive biases in hypothesis formation and evaluation.

2.1.5. Interpretability and justification

Here, we briefly address the distinction between *interpretability*, *explainability*, *justification*, and *explanation*, as used in this article; and as they seem to be used in artificial intelligence.

Lipton [103] provides a taxonomy of the desiderata and methods for interpretable AI. This paper adopts Lipton's assertion that explanation is post-hoc interpretability. I use Biran and Cotton [9]'s definition of *interpretability* of a model as: the degree to which an observer can understand the cause of a decision. 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 or via introspection. I equate 'interpretability' with 'explainability'.

A *justification* explains why a decision is good, but does not necessarily aim to give an explanation of the actual decision-making process [9].

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 explainable 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, although this work is still relevant in many cases.

It is clear that the primary function of explanation is to facilitate *learning* [104,189]. 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 [66] 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 [77,73,69].

Lombrozo [104] 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. Wilkenfeld and Lombrozo [188] go further and show that engaging in explanation but failing to arrive at a correct explanation can improve one's understanding. They describe this as "*explaining for the best inference*", as opposed to the typical model of explanation as "*inference to the best explanation*".

Malle [112, 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 [104] and Wilkenfeld and Lombrozo [188] note that explanations have several functions other than the transfer of knowledge, such as persuasion, learning, 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. For example, it may be better to give a less likely explanation that is more convincing to the explainee if we want them to act in some

positive way. 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 [72, 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* [108,69,77]. 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 [127].

Lipton [102] 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.

Most authors in this area argue that *all* why-questions ask for contrastive explanations, even if the foils are not made explicit [102,77,69,72,110,108], and that people are good at inferring the foil; e.g. from 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. Lipton [102] notes that "central requirement for a sensible contrastive question is that the fact and the foil have a largely similar history, against which the differences stand out. When the histories are disparate, we do not know where to begin to answer the question." This implies that people could use the similarity of the history of facts and possible foils to determine what the explainee's foil truly is.

It is important that the explainee understands the counterfactual case [69]. 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 [77,73,69]. 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.

Van Bouwel and Weber [175] define four types of explanatory question, three of which are contrastive:

Plain fact:	Why does object a have property P?
P-contrast:	Why does object a have property P, rather than property Q?
O-contrast:	Why does object a have property P, while object b has property Q?
T-contrast:	Why does object a have property P at time t, but property Q at time t'?

Van Bouwel and Weber note that differences occur on properties within an object (P-contrast), between objects themselves (O-contrast), and within an object over time (T-contrast). They reject the idea that all 'plain fact' questions have an implicit foil, proposing that plain-fact questions require showing details across a 'non-interrupted' causal chain across time. They argue that plain-fact questions are typically asked due to curiosity, such as desiring to know how certain facts fit into the world, while contrastive questions are typically asked when unexpected events are observed.

Lipton [102] argues that contrastive explanations between a fact P and a foil Q are, in general, easier to derive than 'complete' explanations for plain-fact questions about 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 the P-contrast 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 contrastive 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.

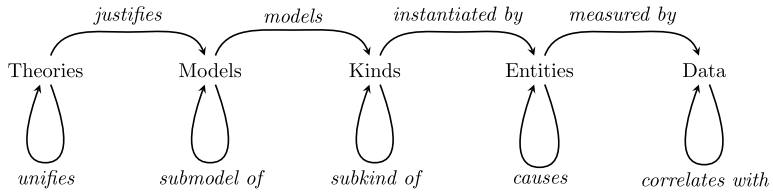

Fig. 3. Overton's five categories and four relations in scientific explanation, reproduced from Overton [139, p. 54, Fig. 3.1].

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 [159]. 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 [60]:

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. An 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 [35], who proposed that people take three *stances* towards objects: *physical*, *design*, and *intention*; and Marr [119], building on earlier work with Poggio [120], who define the *computational*, *representational*, and *hardware* levels of understanding for computational problems.

Kass and Leake [85] define a categorisation of explanations of anomalies into three types: (1) *intentional*; (2) *material*; and (3) *social*. The intentional and material categories correspond roughly to Aristotle's final and material categories, however, the *social* category does not correspond to any particular category in the models of Aristotle, Marr [119], or Dennett [35]. 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 [68] argued for logically deductive models of explanation. Kelley [86] subsequently argued instead that people consider *co-variation* in constructing explanations, and proposed a *statistical* model of explanation. However, while influential, 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 of everyday explanation in most camps [114].

Overton [140,139] defines a model of scientific explanation. In particular, Overton [139] defines the *structure* of explanations. He 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 are shown in Fig. 3.

From these categories, Overton [139] provides a crisp definition of the structure of scientific explanations. He argues that explanations of phenomena at one level must be relative to and refer to at least one other level, and that explanations between two such levels must refer to all intermediate levels. For example, an arthropod (*Entity*) has eight legs (*Data*). Entities of this *Kind* are spiders, according to the *Model* of our *Theory* of arthropods. In this example, the explanation is constructed by appealing to the *Model* of insects, which, in turn, appeals to a particular *Theory* that underlies that *Model*.

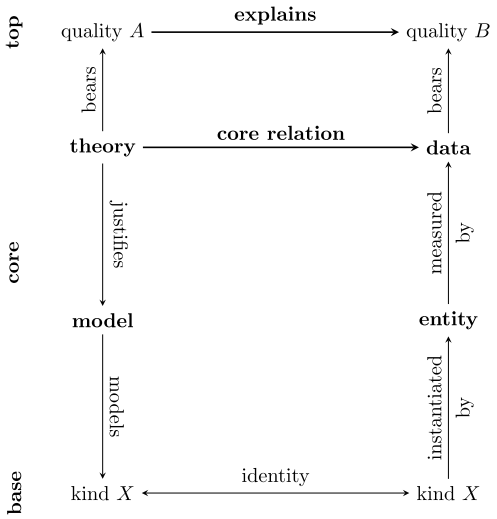

Fig. 4. Overton's general structure of a theory-data explanation, reproduced from Overton [139, p. 54, Fig. 3.2]).

Fig. 4 shows the structure of a *theory-data explanation*, which is the most complex because it has the longest chain of relationships between any two levels.

With respect to social explanation, Malle [112] 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.

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 [112].

2.6. Explanation and XAI

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

2.6.1. Causal attribution is not causal explanation

An important concept is the relationship between cause attribution and explanation. Extracting a causal chain and displaying it to a person is causal attribution, not (necessarily) an explanation. While a person could use such a causal chain to obtain their own explanation, I argue that this does not constitute giving an explanation. In particular, for most AI 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. Even if one agrees with this, 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 often a person may just ask "Why X?", leaving their foil implicit. Eliciting a contrast case from a human observer may be difficult or even infeasible. Lipton [102] 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. As such, the challenge is that the foil needs to be determined. In some applications, the foil could be elicited from the human observer, however, in others, this may not be possible, and therefore, foils may have to be inferred. As noted later in Section 4.6.3, concepts such as *abnormality* could be used to infer likely foils, but techniques for HCI, such as eye gaze [164] and gestures could be used to infer foils in some applications.

It is an opportunity because, as Lipton [102] argues, explaining a contrastive question is often easier than giving a full causal 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: if a person provides a foil, they are implicitly pointing towards the part of the model they do not understand. In Section 4.4, we will see research that outlines how people use contrasts to select explanations that are much simpler than their full counterparts.

Several authors within artificial intelligence flag the importance of contrastive questions. Lim and Dey [100] found via a series of user studies on context-aware applications that "*Why not . . . ?*" questions were common questions that people asked. Further, several authors have looked to answer contrastive questions. For example, Winikoff [190] considers the questions of "*Why don't you believe . . . ?*" and "*Why didn't you do . . . ?*" for BDI programs, or Fox et al. [46] who have similar questions in planning, such as "*Why didn't you do something else (that I would have done)*?". However, most existing work considers contrastive *questions*, but not contrastive *explanations*; that is, finding the differences between the two cases. Providing two complete explanations does not take advantage of contrastive questions. Section 4.4.1 shows that people use the difference between the fact and foil to *focus* explanations on the causes relevant to the question, which makes the explanations more *relevant* to the explainee.

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. Perhaps most why-questions proposed by human observers about such an algorithm would indeed by at the formal level, 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 describes the particular planning algorithm employed; and (4) the final level describes 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) may be more common than in the classification algorithm example. However, questions related to the model are relevant, or why particular actions were taken rather than others, which may depend on the particular optimisation criteria used (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. I argue that these frameworks are useful for analysing the types of questions explanatory agents one may receive. 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 to 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. Explanatory model of self

The work outlined in this section demonstrates that an intelligent agent 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 it's own decision making — a *model of self* — that exists merely for the purpose of explanation. This model may be only an approximation of the original model, but more suitable for explanation.

This idea is not new in XAI. In particular, researchers have investigated machine learning models that are uninterpretable, such as neural nets, and have attempted to extract model approximations using more interpretable model types, such as Bayesian networks [63], decision trees [47], or local approximations [157]. However, my argument here is not only for the purpose of interpretability. Even models considered interpretable, such as decision trees, could be accompanied by another model that is specifically used for explanation. For example, to explain control policies, Hayes and Shah [65] select and annotate particular important state variables and actions that are relevant *for explanation only*. Langley et al. notes that "*An agent must represent content in a way that supports the explanations*" [93, p. 2].

Thus, to generate meaningful and useful explanations of behaviour, models based on the our understanding of explanation must sit alongside and work with the decision-making mechanisms.

2.6.5. Structure of explanation

Related to the 'model of self' is the structure of explanation. Overton's model of scientific explanation [139] defines what I believe to be a solid foundation for the structure of explanation in AI. To provide an explanation along the chain outlined in Fig. 4, one would need an explicit explanatory model (Section 2.6.4) of *each of these different categories* for the given system.

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. Such an approach requires an additional part of the 'model of self' that refers specifically to the learning, not the classification.

Overton's model [139] or one similar to it seems necessary for researchers and practitioners in explainable AI to frame their thoughts and communicate their ideas.

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 [66, 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. For a more detailed survey on this, see McClure [122] and Hilton [70].

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 [66] defines social attribution as *person perception*.

Intentions and *intentionality* is key to the work of Heider [66], and much of the recent work that has followed his — for example, Dennett [35], Malle [112], McClure [122], Boonzaier et al. [10], Kashima et al. [84]. 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 [115] 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 because intentional rule breaking is punished more harshly than unintentional rule breaking. 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 [112]. In other words, folk psychology does not describe how we think; it describes how we think that we think.

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

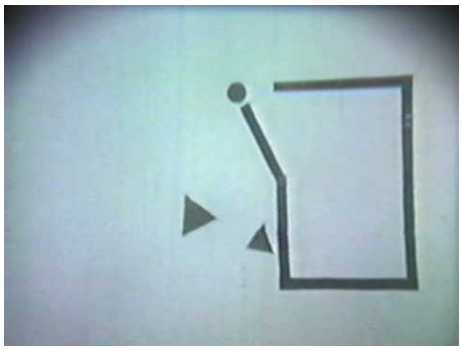

Fig. 5. A screenshot of the video used in Heider and Simmel's seminal study [67].

	Intentional	**Unintentional**
Observable	Actions	Mere behaviours
Unobservable	Intentional thoughts	Experiences

Fig. 6. Malle's classification of types of events, based on the dimensions of intentionality and observability [112, Chapter 3].

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 [66] 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 [67], showed a video containing animated shapes — a small triangle, a large triangle, and a small circle — moving around a screen,[3] and asked experiment participants to observe the video and then describe the behaviour of the shapes. Fig. 5 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 the experiment was to demonstrate that people characterise deliberative behaviour using folk psychology.

Heider [66] 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 [114], incorrectly interpreting Heider's work for decades.

Heider [66] 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 [107] provides the example of Romeo and Juliet, noting that had a wall been placed between them, Romeo would have scaled the wall or knocked in 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 [35,112,122,10,84,108].

Malle and Pearce [118] 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 Fig. 6).

Malle and Pearce [118] performed experiments to confirm this model. As part of these experiments, participants were placed into a room with another participant, 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 observable events, 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

[3] See the video here: https://www.youtube.com/watch?v=VTNmLt7QX8E.

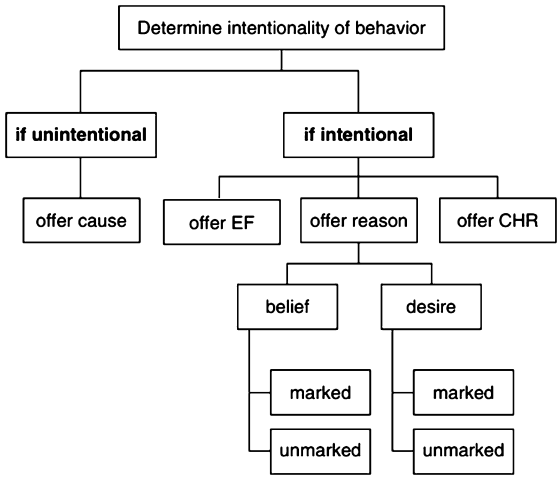

Fig. 7. Malle's conceptual framework for behaviour explanation; reproduced Malle [113, p. 87, Fig. 3.3], adapted from Malle [112, p. 119, Fig. 5.1].

behaviour more *relevant*. Such a model accounts for the *correspondence bias* noted by Gilbert and Malone [51], which is the tendency for people to explain others' behaviours based on traits rather than situational factors, because the situational factors (beliefs, desires) are invisible.

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. [84] 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 [66].

However, it is the extensive body of work from Malle [111–113] that is the most seminal in this space.

3.3.1. Malle's conceptual model for social attribution

Malle [112] 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 Fig. 7. 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.
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 agent's 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 (EF) 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 valuings *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. [167] 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 these individuals/groups, 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 [137] further investigated people's perception of group vs. individual behaviour, focusing on intentionality of explanation. They investigated the relative agency of groups that consist of 'unrelated' individuals acting independently (*aggregate groups*) compared to groups acting together (*jointly acting groups*). In their study, participants were more likely to offer CHR explanations than intention explanations for aggregate groups, and more likely to offer intention explanations than CHR explanations for *jointly* acting groups. For instance, to explain why all people in a department store came to that particular store, participants were 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 came to the same store place, participants were more likely to offer an explanation that the group wanted to spend the day together shopping — a desire. This may demonstrate that people cannot attribute intentional behaviour to the individuals in an aggregate group, so resort to more causal history explanations.

O'Laughlin and Malle's [137] finding about using CHRs to explain aggregate group behaviour is consistent with the earlier work from Kass and Leake [85], whose model of explanation explicitly divided *intentional* explanations from *social explanations*, which are 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 [137] 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 reason 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 their own individual intentions.

3.5. Norms and morals

Norms have been shown to hold a particular place in social attribution. Burguet and Hilton [15] (via Hilton [70]) showed that norms and abnormal behaviour are important in how people ascribe mental states to one another. For example, Hilton [70] 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 [174] 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* [88], which is the effect for people to attribute particular mental states (Theory of Mind) based on moral judgement. Knobe's vignette from his seminal [88] 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 [174] hypothesis that the two existing camps aiming to explain this effect: the *Intuitive Moralist* and *Biased Scientist*, do not account for this. Uttich and Lombrozo 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 [161] 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 [70, 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. [160] 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 XAI

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

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 [67] (Section 3.2) demonstrate unequivocally that people attribute folk psychological concepts such as belief, desire, and intention, to artificial objects. Thus, as argued by de Graaf and Malle [34], it is not a stretch to assert that people will expect explanations using the same conceptual framework used to explain human behaviours.

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 [152], 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 [50].

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 [112], is the most mature and complete model of social attribution to date. His three-layer models 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, and 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, they are also valid explanations for non-anthropomorphic agents. For example, for AI planning agents that optimise some metric, such as cost, the explanation that action a was chosen over action b because it had lower cost is a CHR explanation. 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.

Broekens et al. [12] describe algorithms for automatic generation of explanations for BDI agents. Although their work does not build on Malle's model directly, it shares a similar structure, as noted by the authors, in that their model uses intentions and enabling conditions as explanations. They present three algorithms for explaining behaviour: (a) offering the goal towards which the action contributes; (b) offering the enabling condition of an action; and (c) offering the next action that is to be performed; thus, the explanadum is explained by offering a proximal intention. A set of human behavioural experiments showed that the different explanations are considered better in different circumstances; for example, if only one action is required to achieve the goal, then offering the goal as the explanation is more suitable than offering the other two types of explanation, while if it is part of a longer sequence, also offering a proximal intention is evaluated as being a more valuable explanation. These results reflect those by Malle, but also other results from social and cognitive psychology on the link between goals, proximal intentions, and actions, which are surveyed in Section 4.4.3

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 [11], computational social choice [26], or argumentation [8]. Although this line of work appears to be much less explored than attributions of individual's behaviour, the findings from Kass and Leake [85], Susskind et al., and in particular O'Laughlin and Malle [137] 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 [77,73,69], and violation of normative behaviour is one such abnormality [73]. Thus, 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 [61, p. 54].

Mill [130] 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 the scientifically/philosophically it is "wrong" to select one explanation over another, but offered several cognitive biases 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 [86]. However, many researchers noted that the covariation models failed to explain many observations; for example, people can identify causes between events from a single data point [127,75]; and therefore, more recently, new theories have displaced them, while still acknowledging that the general idea that people using co-variations is valid.

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 the 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. Most of this research shows that people have certain *cognitive biases* that they apply to explanation generation, selection, and evaluation.

4.1. Causal connection, explanation selection, and evaluation

Malle [112] presents a theory of explanation, which breaks the psychological processes used to offer explanations into two distinct groups, outlined in Fig. 8:

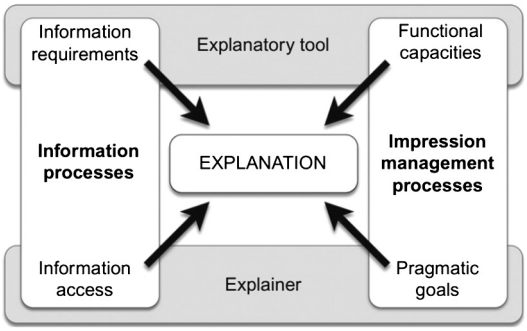

Fig. 8. Malle's process model for behaviour explanation; reproduced from Malle [114, p. 320, Fig. 6.6].

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 [112] 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.

Taking the two dimensions, there are four items:

1. *Information requirements* — what is required to give an adequate explanation; for example, one must knows 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. [117] 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. They hypothesise that 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 to the fact; and *explanation selection* (or *casual selection*), which is the processing of selecting a subset of those causes to provide as an explanation.

This paper separates casual connection into two parts: (1) *abductive reasoning*, the cognitive process in which people try to infer causes that explain events by making assumptions about hypotheses and testing these; and (2) *simulation*, which is the cognitive process of simulating through counterfactuals to derive a good explanation. These processes overlap, but can be somewhat different. For example, the former requires the reasoner to make assumptions and test the validity of observations with respect to these assumptions, while in the latter, the reasoner could have complete knowledge of the causal rules and environment, but use simulation of counterfactual cases to derive an explanation. From the perspective of explainable AI, an explanatory agent explaining its decision would not require abductive reasoning as it is certain of the causes of its decisions. An explanatory agent trying to explain some observed events not under its control, such as the behaviour of another agent, may require abductive reasoning to find a plausible set of causes.

Finally, 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 criterion is that the explanation allows the explainee to understand the cause, however, people's cognitive biases mean that they prefer certain types of explanation over others.

4.2. Causal connection: abductive reasoning

The relationship between explanation and abductive reasoning is introduced in Section 2.1.4. This section surveys work in cognitive science that looks at the process of abduction. Of particular interest to XAI (and artificial intelligence in general) is work demonstrating the link between explanation and learning, but also other processes that people use to simplify

the abductive reasoning process for explanation generation, and to switch modes of reasoning to correspond with types of explanation.

4.2.1. Abductive reasoning and causal types

Rehder [154] looked specifically at *categorical* or *formal* explanations. He presents the *causal model theory*, which states that people infer categories of objects by both their features and the *causal relationships between features*. His experiments show that people categorise objects based their perception that the observed properties were generated by the underlying causal mechanisms. Rehder gives the example that people not only know that birds can fly and bird have wings, but that birds can fly *because* they have wings. In addition, Rehder shows that people use combinations of features as evidence when assigning objects to categories, especially for features that seem incompatible based on the underlying causal mechanisms. For example, when categorising an animal that cannot fly, yet builds a nest in trees, most people would consider it implausible to categorise it as a bird because it is difficult to build a nest in a tree if one cannot fly. However, people are more likely to categorise an animal that does not fly and builds nests on the ground as a bird (e.g. an ostrich or emu), as this is more plausible; even though the first example has more features in common with a bird (building nests in trees).

Rehder [155] extended this work to study how people *generalise* properties based on the explanations received. When his participants were asked to infer their own explanations using abduction, they were more likely to generalise a property from a source object to a target object if they had more features that were similar; e.g. generalise a property from one species of bird to another, but not from a species of bird to a species of plant. However, given an explanation based on features, this relationship is almost completely eliminated: the generalisation was only done if the features detailed in the explanation were shared between the source and target objects; e.g. bird species *A* and mammal *B* both eat the same food, which is explained as the cause for an illness, for example. Thus, the abductive reasoning process used to infer explanations were also used to generalise properties — a parallel seen in machine learning [133].

However, Williams et al. [189] demonstrate that, at least for categorisation in abductive reasoning, the properties of generalisation that support learning can in fact weaken learning by *overgeneralising*. They gave experimental participants a categorisation task to perform by training themselves on exemplars. They asked one group to explain the categorisations as part of the training, and another to just 'think aloud' about their task. The results showed that the explanation group more accurately categorised features that had similar patterns to the training examples, but less accurately categorised exceptional cases and those with unique features. Williams et al. argue that explaining (which forces people to think more systematically about the abduction process) is good for fostering generalisations, but this comes at a cost of over-generalisation.

Chin-Parker and Cantelon [28] provide support for the contrastive account of explanation (see Section 2.3) in categorisation/classification tasks. They hypothesise that *contrast classes* (foils) are key to providing the context to explanation. They distinguish between *prototypical* features of categorisation, which are those features that are typical of a particular category, and *diagnostic* features, which are those features that are relevant for a contrastive explanation. Participants in their study were asked to either describe particular robots or explain why robots were of a particular category, and then follow-up on transfer learning tasks. The results demonstrated that participants in the design group mentioned significantly more features in general, while participants in the explanation group selectively targeted contrastive features. These results provide empirical support for contrastive explanation in category learning.

4.2.2. Background and discounting

Hilton [73] discusses the complementary processes of *backgrounding* and *discounting* that affect the abductive reasoning process. Discounting is when a hypothesis is deemed less likely as a cause because additional contextual 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 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 e.g. the contrastive foil also has this cause.

As noted by Hilton [73], 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.

4.2.3. 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 two most prominent being dependence theories and transference theories.

Lombrozo [107] argues that 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 abductive reasoning 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 [107] 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 were interpreted in a functional manner, counterfactual dependence was important, but physical connections were not. However, when events were interpreted in a mechanistic manner, both counterfactual dependence and physical dependence were both deemed important. This implies that there is a link between functional causation and dependence theories on the one hand, and between mechanistic explanation and transference theories on the other. The participants also rated the functional explanation stronger in the case that the causal dependence was intentional, as opposed to accidental.

Lombrozo [106] studied at the same issue of functional vs. mechanistic explanations for inference in categorisation tasks specifically. She 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 from the first prompt were more likely to think that the flower would bend over, and vice-versa for functional causes. Their findings show that giving explanations influences the inference process, changing 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. In extending work, Lombrozo and Gwynne [109] argue that people generalise better from functional than mechanistic explanations.

4.2.4. Inherent and extrinsic features

Prasada and Dillingham [149] and Prasada [148] discuss how people's abductive reasoning process prioritises certain factors in the formal mode. 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.*" [148, p. 2]

Prasada and Dillingham [149] 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 [139] (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 inherent properties of a thing that are due to its kind, and which they call *principled connections*; and *t-properties*, which are the extrinsic 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, they 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.

Using findings from previous studies, Cimpian and Salomon [30] argue that, when asked to explain a phenomenon, such as a feature of an object, people's cognitive biases make them more likely to use inherent features (k-properties) about the object to explain the phenomenon, rather than extrinsic features (t-properties), such as historical factors. An inherent feature is one that characterises "how an object is constituted" [30, p. 465], and therefore they tend to be stable and enduring features. For example, "spiders have eight legs" is inherent, while "his parents are scared of spiders" is not. Asked to explain why they find spiders scary, people are more likely to refer to the "legginess" of spiders rather than the fact that their parents have arachnophobia, even though studies show that people with arachnophobia are more likely to have family members who find spiders scary [33]. Cimpian and Salomon argue that, even if extrinsic information is known, it is not readily accessible by the *mental shotgun* [82] that people use to retrieve information. For example, looking at spiders, you can see their legs, but not your family's fear of them. Therefore, this leads to people biasing explanations towards inherent features rather than extrinsic. This is similar to the correspondence bias discussed in Section 3.2, in which people are more likely to describe people's behaviour on personality traits rather than beliefs, desires, and intentions, because the latter are not readily accessible while the former are stable and enduring. The bias towards inherence is affected by many factors, such as prior knowledge, cognitive ability, expertise, culture, and age.

4.3. 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 biases used to 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.

4.3.1. Abnormality

Kahneman and Tversky [83] performed seminal work in this field, proposing the *simulation heuristic*. They hypothesise 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 identity primary causes in causal chains using vignettes of a car accident causing the fatality of Mr. Jones, and which had multiple necessary causes, 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 [83] 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.

4.3.2. Temporality

Miller and Gunasegaram [131] 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.

4.3.3. Controllability and intent

Girotto et al. [54] 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.

4.3.4. Social norms

McCloy and Byrne [121] 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. [54], 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 causes 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.

4.4. 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 [70] argues that explanation selection is used for cognitive reasons:

causal chains are often too large to comprehend. He provides an example [70, p. 43, Fig. 7] showing the causal chain for the story of the fatal car accident involving 'Mr. Jones' from Kahneman and Tversky [83]. 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 [172].

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.

4.4.1. 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 [110] 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 [69]:

> "*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 [69, p. 24].

Hesslow [69] 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 [102], building on earlier work in philosophy from Lewis [99], derived similar thoughts to Hesslow [69], 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 [102, p. 256].

From an experimental perspective, Hilton and Slugoski [77] 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 4.4.2). More recent research in cognitive science from Rehder [154,155] supports the theory that people perform causal inference, explanation, and generalisation based on contrastive cases.

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.

4.4.2. Abnormality

Related to the idea of contrastive explanation, Hilton and Slugoski [77] propose the *abnormal conditions model*, based on observations from legal theorists Hart and Honoré [64]. 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 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 to 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 [125], McClure et al. [126], and Hilton et al. [76], and more recently, Samland and Waldmann [161], who show that a variety of non-statistical measures are valid foils.

4.4.3. Intentionality and functionality

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

Hilton et al. [76] investigate the claim from legal theorists Hart and Honoré [64] 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 deliberative 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.

Lombrozo [107] 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.

4.4.4. Necessity, sufficiency and robustness

Several authors [102,107,192] argue that *necessity* and *sufficiency* are strong criteria for preferred explanatory causes. Lipton [102] 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.

Woodward [192] argues that sufficiency is another strong criteria, in that people prefer causes that bring about the effect without any other cause. This should not be confused with sufficiency in the example above, in which either mutation M_1 or M_2 is sufficient in combination with temperature. Woodward's argument applies to uniquely sufficient causes, rather than cases in which there are multiple sufficient causes. For example, if it were found that are third mutation M_3 could cause longer wings irrelevant of the temperature, this would be preferred over temperature plus another mutation. This is related to the notation of *simplicity* discussed in Section 4.5.1.

Finally, several authors [107,192] argue that *robustness* is also a criterion for explanation selection, in which the extend to which a cause C is considered robust is whether the effect E would still have occurred if conditions other than C were somewhat different. Thus, a cause C_1 that holds only in specific situations has less explanatory value than cause C_2, which holds in many other situations.

4.4.5. Responsibility

The notions of *responsibility* and *blame* are relevant to causal selection, in that an event considered 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 'degree of necessity' of causes. An event that is fully responsible outcome for an event is a necessary cause.

Chockler and Halpern [29] modified the structural equation model proposed by Halpern and Pearl [58] (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 (cf. [74,92]), I focus on the model of Chockler and Halpern [29] because, as far I am aware, experimental evaluation of the model shows it to be stronger than existing models [48], 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 [48] 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 [29] 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 off the individual was to the actual number of triangles.

4.4.6. Preconditions, failure, and intentions

An early study into explanation selection in cases of more than one cause was undertaken by Leddo et al. [96]. 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* [173], 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. [96] 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 [106]. 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 [107] argues similarly that this is because failed actions are not goal-directed, because people do not intend to fail. Thus, people prefer *mechanistic* explanations for failed actions, rather than explanations that cite intentions.

McClure and Hilton [123] and McClure et al. [124] 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 [125] 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.

4.5. Explanation evaluation

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

4.5.1. Coherence, simplicity, and generality

Thagard [171] argues that coherence is a primary criterion 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 [151].

Read and Marcus-Newhall [153] tested the hypotheses from Thagard's theory of explanatory coherence [171] that people prefer simpler and more general explanations. Participants were asked to rate the probability and the 'quality' of explanations with different numbers of causes. They 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.

4.5.2. Truth and probability

Probability has two facets in explanation: the probability of the explanation being true; and the *use* of probability in an explanation. Neither has a much importance as one may expect.

The use of statistical relationships to explain events is considered to be unsatisfying on its own. This is because people desire *causes* to explain events, not associative relationships. Josephson and Josephson [81] give the example of a bag full of red balls. When selecting a ball randomly from the bag, it must be red, and one can ask: "Why is this ball red?". The answer that uses the statistical generalisation "Because all balls in the bag are red" is not a good explanation, because it does not explain why that particular ball is red. A better explanation is someone painted it red. However, for the question: "Why did we observe a red ball coming out of the bag", it is a good explanation, because having only red balls in the bag does cause us to select a red one. Josephson and Josephson highlight that the difference between explaining the *fact observed*

(the ball is red) and explaining *the event of observing the fact* (a red ball was selected). To explain instances via statistical generalisations, we need to explain the *causes* of those generalisations too, not the generalisations themselves. If the reader is not convinced, consider my own example: a student coming to their teacher to ask why they only received 50% on an exam. An explanation that most students scored around 50% is not going to satisfy the student. Adding a cause for why most students only scored 50% would be an improvement. Explaining to the student why they specifically received 50% is even better, as it explains the cause of the instance itself.

The truth of likelihood of an explanation is considered an important criterion of a good explanation. However, Hilton [73] shows that the most likely or 'true' cause is not necessarily the *best* explanation. Truth conditions[4] are a necessary but not sufficient criteria for the generation 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 4.4.2), in which a faulty seal was argued to be a better explanation than oxygen in the atmosphere. This is despite the fact 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 [122] also challenges the idea of probability as a criterion 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 [56] (see Section 5.1.1 for a more detailed discussion of this). This is supported by experiments such as Read and Marcus-Newhall [153] cited above, and the work from Tversky and Kahneman [173] on the conjunction fallacy.

Lombrozo [105] notes that the experiments on generality and simplicity performed by Read and Marcus-Newhall [153] 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 how people evaluate explanations, 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 [153], Lombrozo presented experimental participants with information about a patient with several symptoms that could be explained by one cause or several separate causes. In some setups, base rate information about each disease was provided, in which the conjunction of the separate causes was more likely than the single (simpler) cause. Without base-rate information, participants selected the most simple (less likely) explanations. When base-rate information was included, this still occurred, but 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 [150] 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 condition 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.

Kulesza et al. [90] investigate the balance between soundness and completeness of explanation. They investigated explanatory debugging of machine learning algorithms making personalised song recommendations. By using progressively simpler models with less features, they trained a recommender system to give less correct recommendations. Participants were given recommendations for songs on a music social media site, based on their listening history, and were placed into one of several treatments. Participants in each treatment would be given a different combination of soundness and completeness, where soundness means that the explanation is correct and completeness means that all of the underlying causes are identified. For example, one treatment had low soundness but high completeness, while another had medium soundness and medium completeness. Participants were given a list of recommended songs to listen to, along with the (possibly unsound and incomplete) explanations, and were subsequently asked why the song had been recommended. The participants' mental models were measured. The results show that sound and complete models were the best for building a correct mental model, but at the expense of cost/benefit. Complete but unsound explanations improved the participants' mental models more than soundness, and gave a better perception of cost/benefit, but reduced trust. Sound but incomplete explanations were the least preferred, resulting in higher costs and more requests for clarification. Overall, Kulesza et al.

[4] We use the term *truth condition* to refer to facts that are either true or considered likely by the explainee.

concluded that completeness was more important than soundness. From these results, Kulesza et al. [89] list three principles for explainability: (1) *Be sound*; (2) *Be complete*; but (3) *Don't overwhelm*. Clearly, principles 1 and 2 are at odds with principle 3, indicating that careful design must be put into explanatory debugging systems.

4.5.3. Goals and explanatory mode

Vasilyeva et al. [177] 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.

4.6. Cognitive processes and XAI

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

The idea of explanation selection is not new in XAI. Particularly in machine learning, in which models have many features, the problem is salient. Existing work has primarily looked at selecting which features in the model were important for a decision, mostly built on local explanations [158,6,157] or on information gain [90,89]. However, as far as the authors are aware, there are currently no studies that look at the cognitive biases of humans as a way to select explanations from a set of causes.

4.6.1. Abductive reasoning

Using abductive reasoning to generate explanations has a long history in artificial intelligence [97], aiming to solve problems such as fault diagnosis [144], plan/intention recognition [24], and generalisation in learning [133]. Findings from such work has parallels with many of the results from cognitive science/psychology outlined in this section. Leake [95] provides an excellent overview of the challenges of abduction for everyday explanation, and summarises work that addresses these. He notes three of the main tasks that an abductive reasoner must perform are: (1) what to explain about a given situation (determining the question); (2) how to generate explanations (abductive reasoning); and (3) how to evaluate the "best" explanation (explanation selection and evaluation). He stresses that determining the goal of the explanation is key to providing a good explanation; echoing the social scientists' view that the explainee's question is important, and that such questions are typically focused on anomalies or surprising observations.

The work from Rehder [154,155] and Lombrozo [108] show that explanation is good for learning and generalisation. This is interesting and relevant for XAI, because it shows that individual users should require less explanation the more they interact with a system. First, because they will construct a better mental model of the system and be able to generalise its behaviour (effectively learning its model). Second, as they see more cases, they should become less surprised by abnormal phenomena, which as noted in Section 4.4.2, is a primary trigger for requesting explanations. An intelligent agent that presents — unprompted — an explanation alongside every decision, runs a risk of providing explanations that become less needed and more distracting over time.

The work on inherent vs. extrinsic features (Section 4.2.4) is relevant for many AI applications, in particular classification tasks. In preliminary work, Bekele et al. [7] use the inherence bias [30] to explain person identification in images. Their re-identification system is tasked with determining whether two images contain the same person, and uses inherent features such as age, gender, and hair colour, as well as extrinsic features such as clothing or wearing a backpack. Their explanations use the inherence bias with the aim of improving the acceptability of the explanation. In particular, when the image is deemed to be of the same person, extrinsic properties are used, while for different people, intrinsic properties are used. This work is preliminary and has not yet been evaluated, but it is an excellent example of using cognitive biases to improve explanations.

4.6.2. Mutability and computation

Section 4.3 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 to 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.

4.6.3. Abnormality

Abnormality clearly plays a role in explanation and interpretability. For explanation, it serves as a trigger for explanation, and is a useful criterion 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 criterion 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 but valuable.

One important note to make is regarding abnormality and its application to "*non-contrastive*" why-questions. As noted in Section 2.6.2, questions of the form "*Why* P*?*" may have an implicit foil, and determining this can improve explanation. In some cases, normality could 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*?*" [72]. For example, consider the application of assessing the risk of glaucoma [22]. 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.

4.6.4. Intentionality and functionality

The work discussed in Section 4.4.3 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 causes as intentional could be useful. For instance, consider a predictive model in which some features represent that an intentional event has occurred. Prioritising these may lead to more intuitive explanations.

4.6.5. Perspectives and controllability

The finding from Kahneman and Tversky [83] that perspectives change the events people mutate, discussed in Section 4.3, is important in multi-agent contexts. This implies that when explaining a particular agent's decisions or behaviour, the explanatory agent could focus on undoing actions of that particular agent, rather than others. This is also consistent with the research on controllability discussed in Section 4.3, 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 else 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.

4.6.6. Evaluation of explanations

The importance of the research outlined in Section 4.5 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 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.

The selection and evaluation of explanations in artificial intelligence has been studied in some detail, going back to early work on abductive reasoning, in which explanations with structural simplicity, coherence, or minimality are preferred (e.g. [156,97]) and the concept of explanatory power of a set of hypotheses is defined as the set of manifestations those hypotheses account for [1]. Other approaches use probability as the defining factor to determine the most likely explanation (e.g. [59]). In addition to the cognitive biases of people to discount probability, the probabilistic approaches have the problem that such fine-grained probabilities are not always available [95]. These selection mechanisms are context-independent and do not account for the explanations as being relevant to the question nor the explainee.

Leake [94], on the other hand, argues for goal-directed explanations in abductive reasoning that explicitly aim to reduce knowledge gaps; specifically to explain why an observed event is "reasonable" and to help identify faulty reasoning processes that led to it being surprising. He proposes nine *evaluation dimensions* for explanations: timeliness, knowability, distinctiveness, predictive power, causal force, independence, repairability, blockability, and desirability. Some of these cor-

respond to evaluation criteria outlined in Section 4.5; for example, distinctiveness notes that a cause that is surprising is of good explanatory value, which equates to the criteria of abnormality.

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 [72].

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 [72] 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*, which 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*" [72, p. 66].

The conversational model argues that good social explanations must be *relevant*. This means that they must answer the 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 [56], including being relevant to the explainee themselves, and what they already know.

Hilton [72] 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* [56]: (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 (epistemic explanation selection).

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 [72].

5.1.1. Logic and conversation

Grice's maxims [56] (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*" [56, p. 45].

For this, Grice [56] 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 [56] 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 [72] 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 [69, 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 4.4.1, we saw evidence to suggest that the difference between the fact and foil for contrastive why-questions 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. [165] 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 a 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 [80] and Hilton [73] 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 [16] 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 [59] present an elegant formal model of explanation selection based on epistemic relevance. This model extends their work on structural causal models [59], 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 [58] contain variables and functions between these variables. A *situation* is a unique assignment from variables to values. Halpern and Pearl [59] then define an *epistemic state* as a set of situations, one for each possible situation that the explainee considers possible. Explaining the causes of an event then becomes providing the values for those variables that remove some situations from the epistemic state such that 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 [38] 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 [169] 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. [170] 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. Argumentation and explanation

Antaki and Leudar [3] extend Hilton's conversational model [72] from 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 [3, 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.

5.1.4. Linguistic structure

Malle et al. [116] 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*" [116].[5]

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 [112, 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.

[5] Malle [112, Chapter 4] also briefly discusses *valuings* as markers, such as "*She likes*", but notes that these are rarely dropped in reasons.

Such constructs aim not to provide different reasons, but merely allow the speaker to form impressions about themselves and the actor.

5.2. Explanatory dialogue

If we accept the 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 [176] to definitions based on transfer of understanding [179]. However, the most relevant for the problem of explanation in AI is a body of work lead largely by Walton.

Walton [180] proposed a dialectical theory of explanation, putting forward similar ideas to that of Antaki and Leudar [3] 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 [182] proposes a formal dialogue model called *CE*, based on an earlier persuasion dialogue [184], which defines the conditions on how a explanatory 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 [182], Walton [183] 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 [162] 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* [181] 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 [4] formalise Walton's work on explanation dialogue [183], grounding it in a well-known argumentation framework [147]. 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 [3]. 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 XAI

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

5.3.1. Conversational model

The conversational model of explanation according to Hilton [72], and its subsequent extension by Antaki and Leudar [3] to consider argumentation, are appealing and useful models for explanation in AI. In particular, they are appealing because of its generality — they 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 [72], Antaki and Leudar [3], and Walton [183] are all 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 maxim 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. For example, what should an explanatory agent do if the explainee does not accept a selected explanation?

5.3.2. Dialogue

Walton's explanation dialogues [180,182,183], which build on well-accepted models from argumentation, are closer to the notion of computational models than that of Hilton [72] or Antaki and Leudar [3]. 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.

Arioua and Croitoru's *formal* model for explanation [4] not only brings us one step closer to a computational model, but also nicely brings together the models of Hilton [72] and Antaki and Leudar [3] for allowing arguments over claims in explanations. Such formal models of explanation could work together with concepts such as *conversation policies* [55] to implement explanations.

The idea of interactive dialogue XAI is not new. In particular, a body of work by Cawsey [17–19] describes EDGE: a system that generates natural-language dialogues for explaining complex principles. Cawsey's work was novel because it was the first to investigate discourse within an explanation, rather than discourse more generally. Due to the complexity of explanation, Cawsey advocates *context-specific*, *incremental* explanation, interleaving planning and execution of an explanation dialogue. EDGE separates content planning (what to say) from dialogue planning (organisation of the interaction). Interruptions attract their own sub-dialog. The flow of the dialogue is context dependent, in which context is given by: (1) the current state of the discourse relative to the goal/sub-goal hierarchy; (2) the current *focus* of the explanation, such as which components of a device are currently under discussion; and (3) assumptions about the user's knowledge. Both the content and dialogue are influenced by the context. The dialogue is planned using a rule-based system that break explanatory goals into sub-goals and utterances. Evaluation of EDGE [19] is anecdotal, based on a small set of people, and with no formal evaluation or comparison.

At a similar time, Moore and Paris [134] devised a system for explanatory text generation within dialogues that also considers context. They explicitly reject the notion that *schemata* can be used to generate explanations, because they are too rigid and lack the intentional structure to recover from failures or misunderstandings in the dialogue. Like Cawsey's EDGE system, Moore and Paris explicitly represent the user's knowledge, and plan dialogues incrementally. The two primary differences from EDGE is that Moore and Paris's system explicitly models the effects that utterances can have on the hearer's mental state, providing flexibility that allows recovery from failure and misunderstanding; and that the EDGE system follows an extended explanatory plan, including probing questions, which are deemed less appropriate in Moore and Paris's application area of advisory dialogues. The focus of Cawsey's and Moore and Paris's work are in applications such as intelligent tutoring, rather than on AI that explains itself, but many of the lessons and ideas generalise.

EDGE and other related research on interactive explanation considers only verbal dialogue. As noted above, abstract models of dialogue such as those proposed by Walton [183] may serve as a good starting point for multi-model interactive explanations.

5.3.3. Theory of mind

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 *Theory of Mind* 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 and forms part of the explanatory context. 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. Halpern and Pearl [59] already considers a simplified idea of this in their model of explanation, but other work on epistemic reasoning and planning [42,135] and planning for interactive dialogue [143] can play a part here. These techniques will be made more powerful if they are aligned with user modelling techniques used in HCI [44].

While the idea of Theory of Mind in AI is not new; see for example [178,37]; it's application to explanation has not been adequately explored. Early work on XAI took the idea of dialogue and user modelling seriously. For example, Cawsey's EDGE system, described in Section 5.3.2, contains a specific user model to provide better context for interactive explanations [20]. Cawsey argues that the user model must be integrated closely with explanation model to provide more natural dialogue. The EDGE user model consists of two parts: (1) the *knowledge* that the user has about a phenomenon; and (2) their 'level of expertise'; both of which can be updated during the dialogue. EDGE uses dialogues questions to build a user model, either explicitly, using questions such as "Do you known X?" or "What is the value of Y?", or implicitly, such as when a user asks for clarification. EDGE tries to guess other indirect knowledge using logical inference from this direct knowledge. This knowledge is then used to tailor explanation to the specific person, which is an example of using epistemic relevance to select explanations. Cawsey was not the first to consider user knowledge; for example, Weiner's BLAH system [185] for incremental explanation also had a simple user model for knowledge that is used to tailor explanation, and Weiner refers to Grice's maxim of quality to justify this.

More recently, Chakraborti et al. [21] discuss preliminary work in this area for explaining plans. Their problem definition consists of two planning models: the explainer and the explainee; and the task is to align the two models by minimising

some criteria; for example, the number of changes. This is an example of using epistemic relevance to tailor an explanation. Chakraborti et al. class this as contrastive explanation, because the explanation contrasts two models. However, this is not the same use of the term 'contrastive' as used in social science literature (see Section 2.3), in which the contrast is an explicit foil provided by the explainee as part of a question.

5.3.4. 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 [38], 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.

5.3.5. 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.

5.3.6. Social and interactive explanation

The recent surge in explainable AI has not (yet) truly adopted the concept socially-interactive explanation, at least, relative to the first wave of explainable AI systems such as that by Cawsey [20] and Moore and Paris [134]. I hypothesise that this is largely due to the nature of the task being explained. Most recent research is concerned with explainable machine learning, whereas early work explained symbolic models such as expert systems and logic programs. This influences the research in two ways: (1) recent research focuses on how to abstract and simplify uninterpretable models such as neural nets, whereas symbolic approaches are relatively more interpretable and need less abstraction in general; and (2) an interactive explanation is a goal-based endeavour, which lends itself more naturally to symbolic approaches. Given that early work on XAI was to explain symbolic approaches, the authors of such work would have more intuitively seen the link to interaction. Despite this, others in the AI community have recently re-discovered the importance of social interaction for explanation; for example, [186,163], and have noted that this is a problem that requires collaboration with HCI researchers.

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 provided some insight into how this work can be used in explainable AI.

In particular, we should take the four major findings noted in the introduction into account in our explainable AI models: (1) why-questions are contrastive; (2) explanations are selected (in a biased manner); (3) explanations are social; and (4) probabilities are not as important as causal links. I acknowledge that incorporating these ideas are not 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.

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 philosophy, psychology, cognitive science, and human–computer interaction. Already, projects of this type are underway, with impressive results; for example, see [91,89,157].

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 several others for critical input of an earlier draft: Natasha Goss, Michael Winikoff, Gary Klein, Robert Hoffman, and the anonymous reviewers; and Darryn Reid for his discussions on the link between self, trust, and explanation.

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 a Sponsored Research Collaboration grant from the Commonwealth of Australia Defence Science and Technology Group and the Defence Science Institute, an initiative of the State Government of Victoria.

References

[1] D. Allemang, M.C. Tanner, T. Bylander, J.R. Josephson, Computational complexity of hypothesis assembly, in: IJCAI, vol. 87, 1987, pp. 1112–1117.
[2] J. Angwin, J. Larson, S. Mattu, L. Kirchner, Machine bias, ProPublica (23 May 2016).
[3] C. Antaki, I. Leudar, Explaining in conversation: towards an argument model, Eur. J. Soc. Psychol. 22 (2) (1992) 181–194.
[4] A. Arioua, M. Croitoru, Formalizing explanatory dialogues, in: International Conference on Scalable Uncertainty Management, Springer, 2015, pp. 282–297.
[5] J.L. Aronson, On the grammar of 'cause', Synthese 22 (3) (1971) 414–430.
[6] D. Baehrens, T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen, K.-R. MÃžller, How to explain individual classification decisions, J. Mach. Learn. Res. 11 (Jun) (2010) 1803–1831.
[7] E. Bekele, W.E. Lawson, Z. Horne, S. Khemlani, Human-level explanatory biases for person re-identification, in: HRI Workshop on Explainable Robotic Systems, 2018.
[8] P. Besnard, A. Hunter, Elements of Argumentation, vol. 47, MIT Press, Cambridge, 2008.
[9] O. Biran, C. Cotton, Explanation and justification in machine learning: a survey, in: IJCAI 2017 Workshop on Explainable Artificial Intelligence (XAI), 2017, pp. 8–13.
[10] A. Boonzaier, J. McClure, R.M. Sutton, Distinguishing the effects of beliefs and preconditions: the folk psychology of goals and actions, Eur. J. Soc. Psychol. 35 (6) (2005) 725–740.
[11] R.I. Brafman, C. Domshlak, From one to many: planning for loosely coupled multi-agent systems, in: International Conference on Automated Planning and Scheduling, 2008, pp. 28–35.
[12] J. Broekens, M. Harbers, K. Hindriks, K. Van Den Bosch, C. Jonker, J.-J. Meyer, Do you get it? User-evaluated explainable BDI agents, in: German Conference on Multiagent System Technologies, Springer, 2010, pp. 28–39.
[13] S. Bromberger, Why-questions, in: R.G. Colodny (Ed.), Mind and Cosmos: Essays in Contemporary Science and Philosophy, Pittsburgh University Press, Pittsburgh, 1966, pp. 68–111.
[14] B. Buchanan, E. Shortliffe, Rule-based Expert Systems: the MYCIN Experiments the Stanford Heuristic Programming Project, Addison-Wesley, 1984.
[15] A. Burguet, D. Hilton, Effets de contexte sur l'explication causale, in: M. Bromberg, A. Trognon (Eds.), Psychologie Sociale et Communication, Dunod, Paris, 2004, pp. 219–228.
[16] R.M. Byrne, The construction of explanations, in: AI and Cognitive Science'90, Springer, 1991, pp. 337–351.
[17] A. Cawsey, Generating interactive explanations, in: AAAI, 1991, pp. 86–91.
[18] A. Cawsey, Explanation and Interaction: The Computer Generation of Explanatory Dialogues, MIT Press, 1992.
[19] A. Cawsey, Planning interactive explanations, Int. J. Man-Mach. Stud. 38 (2) (1993) 169–199.
[20] A. Cawsey, User modelling in interactive explanations, User Model. User-Adapt. Interact. 3 (1993) 221–247.
[21] T. Chakraborti, S. Sreedharan, Y. Zhang, S. Kambhampati, Plan explanations as model reconciliation: moving beyond explanation as soliloquy, in: Proceedings of IJCAI, 2017, https://www.ijcai.org/proceedings/2017/0023.pdf.
[22] 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 Trans. Biomed. Eng. 49 (9) (2002) 963–974.
[23] B. Chandrasekaran, M.C. Tanner, J.R. Josephson, Explaining control strategies in problem solving, IEEE Expert 4 (1) (1989) 9–15.
[24] E. Charniak, R. Goldman, A probabilistic model of plan recognition, in: Proceedings of the Ninth National Conference on Artificial Intelligence—Volume 1, AAAI Press, 1991, pp. 160–165.
[25] 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.
[26] Y. Chevaleyre, 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, 2007, pp. 51–69.
[27] 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, Cogn. Process. 11 (3) (2010) 227–249.
[28] S. Chin-Parker, J. Cantelon, Contrastive constraints guide explanation-based category learning, Cogn. Sci. 41 (6) (2017) 1645–1655.
[29] H. Chockler, J.Y. Halpern, Responsibility and blame: a structural-model approach, J. Artif. Intell. Res. 22 (2004) 93–115.
[30] A. Cimpian, E. Salomon, The inherence heuristic: an intuitive means of making sense of the world, and a potential precursor to psychological essentialism, Behav. Brain Sci. 37 (5) (2014) 461–480.
[31] 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.
[32] 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.
[33] G.C. Davey, Characteristics of individuals with fear of spiders, Anxiety Res. 4 (4) (1991) 299–314.
[34] M.M. de Graaf, B.F. Malle, How people explain action (and autonomous intelligent systems should too), in: AAAI Fall Symposium on Artificial Intelligence for Human–Robot Interaction, 2017.
[35] D.C. Dennett, The Intentional Stance, MIT Press, 1989.
[36] D.C. Dennett, From Bacteria to Bach and Back: The Evolution of Minds, WW Norton & Company, 2017.
[37] F. Dignum, R. Prada, G.J. Hofstede, From autistic to social agents, in: Proceedings of the 2014 International Conference on Autonomous Agents and Multi-Agent Systems, IFAAMAS, 2014, pp. 1161–1164.
[38] D.H. Dodd, J.M. Bradshaw, Leading questions and memory: pragmatic constraints, J. Mem. Lang. 19 (6) (1980) 695.
[39] P. Dowe, Wesley Salmon's process theory of causality and the conserved quantity theory, Philos. Sci. 59 (2) (1992) 195–216.
[40] T. Eiter, T. Lukasiewicz, Complexity results for structure-based causality, Artif. Intell. 142 (1) (2002) 53–89.
[41] T. Eiter, T. Lukasiewicz, Causes and explanations in the structural-model approach: tractable cases, Artif. Intell. 170 (6–7) (2006) 542–580.
[42] R. Fagin, J. Halpern, Y. Moses, M. Vardi, Reasoning About Knowledge, Vol. 4, MIT Press, Cambridge, 1995.
[43] D. Fair, Causation and the flow of energy, Erkenntnis 14 (3) (1979) 219–250.
[44] G. Fischer, User modeling in human–computer interaction, User Model. User-Adapt. Interact. 11 (1–2) (2001) 65–86.
[45] J. Fox, D. Glasspool, D. Grecu, S. Modgil, M. South, V. Patkar, Argumentation-based inference and decision making—a medical perspective, IEEE Intell. Syst. 22 (6) (2007) 34–41.
[46] M. Fox, D. Long, D. Magazzeni, Explainable planning, in: IJCAI 2017 Workshop on Explainable Artificial Intelligence (XAI), 2017, https://arxiv.org/pdf/1709.10256.
[47] N. Frosst, G. Hinton, Distilling a neural network into a soft decision tree, arXiv e-prints 1711.09784, https://arxiv.org/abs/1711.09784.
[48] T. Gerstenberg, D.A. Lagnado, Spreading the blame: the allocation of responsibility amongst multiple agents, Cognition 115 (1) (2010) 166–171.
[49] T. Gerstenberg, M.F. Peterson, N.D. Goodman, D.A. Lagnado, J.B. Tenenbaum, Eye-tracking causality, Psychol. Sci. 28 (12) (2017) 1731–1744.
[50] M. Ghallab, D. Nau, P. Traverso, Automated Planning: Theory and Practice, Elsevier, 2004.

[51] D.T. Gilbert, P.S. Malone, The correspondence bias, Psychol. Bull. 117 (1) (1995) 21.
[52] C. Ginet, In defense of a non-causal account of reasons explanations, J. Ethics 12 (3–4) (2008) 229–237.
[53] L. Giordano, C. Schwind, Conditional logic of actions and causation, Artif. Intell. 157 (1–2) (2004) 239–279.
[54] V. Girotto, P. Legrenzi, A. Rizzo, Event controllability in counterfactual thinking, Acta Psychol. 78 (1) (1991) 111–133.
[55] M. Greaves, H. Holmback, J. Bradshaw, What is a conversation policy? in: Issues in Agent Communication, Springer, 2000, pp. 118–131.
[56] H.P. Grice, Logic and conversation, in: Syntax and Semantics 3: Speech Arts, Academic Press, New York, 1975, pp. 41–58.
[57] J.Y. Halpern, Axiomatizing causal reasoning, J. Artif. Intell. Res. 12 (2000) 317–337.
[58] J.Y. Halpern, J. Pearl, Causes and explanations: a structural-model approach. Part I: causes, Br. J. Philos. Sci. 56 (4) (2005) 843–887.
[59] J.Y. Halpern, J. Pearl, Causes and explanations: a structural-model approach. Part II: explanations, Br. J. Philos. Sci. 56 (4) (2005) 889–911.
[60] R.J. Hankinson, Cause and Explanation in Ancient Greek Thought, Oxford University Press, 2001.
[61] N.R. Hanson, Patterns of Discovery: An Inquiry Into the Conceptual Foundations of Science, CUP Archive, 1965.
[62] G.H. Harman, The inference to the best explanation, Philos. Rev. 74 (1) (1965) 88–95.
[63] M. Harradon, J. Druce, B. Ruttenberg, Causal learning and explanation of deep neural networks via autoencoded activations, arXiv e-prints 1802.00541, https://arxiv.org/abs/1802.00541.
[64] H.L.A. Hart, T. Honoré, Causation in the Law, OUP, Oxford, 1985.
[65] 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), p. 2017.
[66] F. Heider, The Psychology of Interpersonal Relations, Wiley, New York, 1958.
[67] F. Heider, M. Simmel, An experimental study of apparent behavior, Am. J. Psychol. 57 (2) (1944) 243–259.
[68] C.G. Hempel, P. Oppenheim, Studies in the logic of explanation, Philos. Sci. 15 (2) (1948) 135–175.
[69] G. Hesslow, The problem of causal selection, in: Contemporary Science and Natural Explanation: Commonsense Conceptions of Causality, 1988, pp. 11–32.
[70] D. Hilton, Social attribution and explanation, in: Oxford Handbook of Causal Reasoning, Oxford University Press, 2017, pp. 645–676.
[71] D.J. Hilton, Logic and causal attribution, in: Contemporary Science and Natural Explanation: Commonsense Conceptions of Causality, New York University Press, 1988, pp. 33–65.
[72] D.J. Hilton, Conversational processes and causal explanation, Psychol. Bull. 107 (1) (1990) 65–81.
[73] D.J. Hilton, Mental models and causal explanation: judgements of probable cause and explanatory relevance, Think. Reasoning 2 (4) (1996) 273–308.
[74] 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, Routledge, London, 2005, pp. 44–60.
[75] D.J. Hilton, J. McClure, R.M. Sutton, Selecting explanations from causal chains: do statistical principles explain preferences for voluntary causes? Eur. J. Soc. Psychol. 40 (3) (2010) 383–400.
[76] D.J. Hilton, J.L. McClure, R. Ben Slugoski, 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.
[77] D.J. Hilton, B.R. Slugoski, Knowledge-based causal attribution: the abnormal conditions focus model, Psychol. Rev. 93 (1) (1986) 75.
[78] R.R. Hoffman, G. Klein, Explaining explanation, part 1: theoretical foundations, IEEE Intell. Syst. 32 (3) (2017) 68–73.
[79] D. Hume, An Enquiry Concerning Human Understanding: A Critical Edition, vol. 3, Oxford University Press, 2000.
[80] J.M. Jaspars, D.J. Hilton, Mental models of causal reasoning, in: The Social Psychology of Knowledge, Cambridge University Press, 1988, pp. 335–358.
[81] J.R. Josephson, S.G. Josephson, Abductive Inference: Computation, Philosophy, Technology, Cambridge University Press, 1996.
[82] D. Kahneman, Thinking, Fast and Slow, Macmillan, 2011.
[83] D. Kahneman, A. Tversky, The simulation heuristic, in: P.S.D. Kahneman, A. Tversky (Eds.), Judgment Under Uncertainty: Heuristics and Biases, Cambridge University Press, New York, 1982.
[84] Y. Kashima, A. McKintyre, P. Clifford, The category of the mind: folk psychology of belief, desire, and intention, Asian J. Social Psychol. 1 (3) (1998) 289–313.
[85] A. Kass, D. Leake, Types of Explanations, Tech. Rep. ADA183253, DTIC Document, 1987.
[86] H.H. Kelley, Attribution Theory in Social Psychology, in: Nebraska Symposium on Motivation, University of Nebraska Press, 1967, pp. 192–238.
[87] H.H. Kelley, Causal Schemata and the Attribution Process, General Learning Press, Morristown, NJ, 1972.
[88] J. Knobe, Intentional action and side effects in ordinary language, Analysis 63 (279) (2003) 190–194.
[89] 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, 2015, pp. 126–137.
[90] 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: 2013 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC), IEEE, 2013, pp. 3–10.
[91] 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 Trans. Interact. Intell. Syst. (TiiS) 1 (1) (2011) 2.
[92] D.A. Lagnado, S. Channon, Judgments of cause and blame: the effects of intentionality and foreseeability, Cognition 108 (3) (2008) 754–770.
[93] 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.
[94] D.B. Leake, Goal-based explanation evaluation, Cogn. Sci. 15 (4) (1991) 509–545.
[95] D.B. Leake, Abduction, experience, and goals: a model of everyday abductive explanation, J. Exp. Theor. Artif. Intell. 7 (4) (1995) 407–428.
[96] J. Leddo, R.P. Abelson, P.H. Gross, Conjunctive explanations: when two reasons are better than one, J. Pers. Soc. Psychol. 47 (5) (1984) 933.
[97] H.J. Levesque, A knowledge-level account of abduction, in: IJCAI, 1989, pp. 1061–1067.
[98] D. Lewis Causation, J. Philos. 70 (17) (1974) 556–567.
[99] D. Lewis, Causal explanation, Philos. Pap. 2 (1986) 214–240.
[100] B.Y. Lim, A.K. Dey, Assessing demand for intelligibility in context-aware applications, in: Proceedings of the 11th International Conference on Ubiquitous Computing, ACM, 2009, pp. 195–204.
[101] 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, Proc. Human Factors Ergonom. Soc. Annual Meeting 50 (23) (2006) 2482–2486.
[102] P. Lipton, Contrastive explanation, R. Inst. Philos. Suppl. 27 (1990) 247–266.
[103] Z.C. Lipton, The mythos of model interpretability, arXiv preprint arXiv:1606.03490.
[104] T. Lombrozo, The structure and function of explanations, Trends Cogn. Sci. 10 (10) (2006) 464–470.
[105] T. Lombrozo, Simplicity and probability in causal explanation, Cogn. Psychol. 55 (3) (2007) 232–257.
[106] T. Lombrozo, Explanation and categorization: how "why?" informs "what?", Cognition 110 (2) (2009) 248–253.
[107] T. Lombrozo, Causal-explanatory pluralism: how intentions, functions, and mechanisms influence causal ascriptions, Cogn. Psychol. 61 (4) (2010) 303–332.
[108] T. Lombrozo, Explanation and abductive inference, in: Oxford Handbook of Thinking and Reasoning, 2012, pp. 260–276.

[109] T. Lombrozo, N.Z. Gwynne, Explanation and inference: mechanistic and functional explanations guide property generalization, Front. Human Neurosci. 8 (2014) 700.
[110] J.L. Mackie, The Cement of the Universe, Oxford, 1980.
[111] B.F. Malle, How people explain behavior: a new theoretical framework, Personal. Soc. Psychol. Rev. 3 (1) (1999) 23–48.
[112] B.F. Malle, How the Mind Explains Behavior: Folk Explanations, Meaning, and Social Interaction, MIT Press, 2004.
[113] B.F. Malle, Attribution theories: how people make sense of behavior, in: Theories in Social Psychology, 2011, pp. 72–95.
[114] B.F. Malle, Time to give up the dogmas of attribution: an alternative theory of behavior explanation, Adv. Exp. Soc. Psychol. 44 (1) (2011) 297–311.
[115] B.F. Malle, J. Knobe, The folk concept of intentionality, J. Exp. Soc. Psychol. 33 (2) (1997) 101–121.
[116] 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, J. Pers. Soc. Psychol. 79 (3) (2000) 309.
[117] B.F. Malle, J.M. Knobe, S.E. Nelson, Actor-observer asymmetries in explanations of behavior: new answers to an old question, J. Pers. Soc. Psychol. 93 (4) (2007) 491.
[118] B.F. Malle, G.E. Pearce, Attention to behavioral events during interaction: two actor-observer gaps and three attempts to close them, J. Pers. Soc. Psychol. 81 (2) (2001) 278–294.
[119] D. Marr, Vision: A Computational Investigation Into the Human Representation and Processing of Visual Information, Inc., New York, NY, 1982.
[120] D. Marr, T. Poggio, From Understanding Computation to Understanding Neural Circuitry, AI Memos AIM-357, MIT, 1976.
[121] R. McCloy, R.M. Byrne, Counterfactual thinking about controllable events, Mem. Cogn. 28 (6) (2000) 1071–1078.
[122] J. McClure, Goal-based explanations of actions and outcomes, Eur. Rev. Soc. Psychol. 12 (1) (2002) 201–235.
[123] J. McClure, D. Hilton, For you can't always get what you want: when preconditions are better explanations than goals, Br. J. Soc. Psychol. 36 (2) (1997) 223–240.
[124] J. McClure, D. Hilton, J. Cowan, L. Ishida, M. Wilson, When rich or poor people buy expensive objects: is the question how or why? J. Lang. Soc. Psychol. 20 (2001) 229–257.
[125] J. McClure, D.J. Hilton, Are goals or preconditions better explanations? It depends on the question, Eur. J. Soc. Psychol. 28 (6) (1998) 897–911.
[126] 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, 2003, p. 306.
[127] A.L. McGill, J.G. Klein, Contrastive and counterfactual reasoning in causal judgment, J. Pers. Soc. Psychol. 64 (6) (1993) 897.
[128] P. Menzies, H. Price, Causation as a secondary quality, Br. J. Philos. Sci. 44 (2) (1993) 187–203.
[129] 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, Hum. Factors 58 (3) (2016) 401–415.
[130] J.S. Mill, A System of Logic: The Collected Works of John Stuart Mill, vol. III, 1973.
[131] D.T. Miller, S. Gunasegaram, Temporal order and the perceived mutability of events: implications for blame assignment, J. Pers. Soc. Psychol. 59 (6) (1990) 1111.
[132] T. Miller, P. Howe, L. Sonenberg, Explainable AI: beware of inmates running the asylum, in: IJCAI 2017 Workshop on Explainable Artificial Intelligence (XAI), 2017, pp. 36–42, http://people.eng.unimelb.edu.au/tmiller/pubs/explanation-inmates.pdf.
[133] T.M. Mitchell, R.M. Keller, S.T. Kedar-Cabelli, Explanation-based generalization: a unifying view, Mach. Learn. 1 (1) (1986) 47–80.
[134] J.D. Moore, C.L. Paris, Planning text for advisory dialogues: capturing intentional and rhetorical information, Comput. Linguist. 19 (4) (1993) 651–694.
[135] 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, 2015, pp. 1–8.
[136] G. Nott, 'Explainable Artificial Intelligence': cracking open the black box of AI, Computer World, https://www.computerworld.com.au/article/617359/.
[137] M.J. O'Laughlin, B.F. Malle, How people explain actions performed by groups and individuals, J. Pers. Soc. Psychol. 82 (1) (2002) 33.
[138] J. Overton, Scientific explanation and computation, in: D.B.L. Thomas Roth-Berghofer, Nava Tintarev (Eds.), Proceedings of the 6th International Explanation-Aware Computing (ExaCt) Workshop, 2011, pp. 41–50.
[139] J.A. Overton, Explanation in Science, Ph.D. thesis, The University of Western, Ontario, 2012.
[140] J.A. Overton, "Explain" in scientific discourse, Synthese 190 (8) (2013) 1383–1405.
[141] J. Pearl, D. Mackenzie, The Book of Why: The New Science of Cause and Effect, Hachette, UK, 2018.
[142] C.S. Peirce, Harvard lectures on pragmatism, in: Collected Papers, vol. 5, 1903.
[143] 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.
[144] D. Poole, Normality and faults in logic-based diagnosis, in: IJCAI, vol. 89, 1989, pp. 1304–1310.
[145] H.E. Pople, On the mechanization of abductive logic, in: IJCAI, vol. 73, 1973, pp. 147–152.
[146] K. Popper, The Logic of Scientific Discovery, Routledge, 2005.
[147] H. Prakken, Formal systems for persuasion dialogue, Knowl. Eng. Rev. 21 (02) (2006) 163–188.
[148] S. Prasada, The scope of formal explanation, Psychon. Bull. Rev. (2017) 1–10.
[149] S. Prasada, E.M. Dillingham, Principled and statistical connections in common sense conception, Cognition 99 (1) (2006) 73–112.
[150] J. Preston, N. Epley, Explanations versus applications: the explanatory power of valuable beliefs, Psychol. Sci. 16 (10) (2005) 826–832.
[151] M. Ranney, P. Thagard, Explanatory coherence and belief revision in naive physics, in: Proceedings of the Tenth Annual Conference of the Cognitive Science Society, 1988, pp. 426–432.
[152] A.S. Rao, M.P. Georgeff, BDI agents: from theory to practice, in: ICMAS, vol. 95, 1995, pp. 312–319.
[153] S.J. Read, A. Marcus-Newhall, Explanatory coherence in social explanations: a parallel distributed processing account, J. Pers. Soc. Psychol. 65 (3) (1993) 429.
[154] B. Rehder, A causal-model theory of conceptual representation and categorization, J. Exp. Psychol. Learn. Mem. Cogn. 29 (6) (2003) 1141.
[155] B. Rehder, When similarity and causality compete in category-based property generalization, Mem. Cogn. 34 (1) (2006) 3–16.
[156] R. Reiter, A theory of diagnosis from first principles, Artif. Intell. 32 (1) (1987) 57–95.
[157] 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, 2016, pp. 1135–1144.
[158] M. Robnik-Šikonja, I. Kononenko, Explaining classifications for individual instances, IEEE Trans. Knowl. Data Eng. 20 (5) (2008) 589–600.
[159] W.C. Salmon, Four Decades of Scientific Explanation, University of Pittsburgh Press, 2006.
[160] J. Samland, M. Josephs, M.R. Waldmann, H. Rakoczy, The role of prescriptive norms and knowledge in children's and adults' causal selection, J. Exp. Psychol. Gen. 145 (2) (2016) 125.
[161] 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, 2014, pp. 1359–1364.
[162] 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, W. H. Winston & Sons, Washington, 1972, pp. 31–39.
[163] Z. Shams, M. de Vos, N. Oren, J. Padget, Normative practical reasoning via argumentation and dialogue, in: Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI-16), AAAI Press, 2016.

[164] R. Singh, T. Miller, J. Newn, L. Sonenberg, E. Velloso, F. Vetere, Combining planning with gaze for online human intention recognition, in: Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems, 2018.
[165] B.R. Slugoski, M. Lalljee, R. Lamb, G.P. Ginsburg, Attribution in conversational context: effect of mutual knowledge on explanation-giving, Eur. J. Soc. Psychol. 23 (3) (1993) 219–238.
[166] K. Stubbs, P. Hinds, D. Wettergreen, Autonomy and common ground in human–robot interaction: a field study, IEEE Intell. Syst. 22 (2) (2007) 42–50.
[167] J. Susskind, K. Maurer, V. Thakkar, D.L. Hamilton, J.W. Sherman, Perceiving individuals and groups: expectancies, dispositional inferences, and causal attributions, J. Pers. Soc. Psychol. 76 (2) (1999) 181.
[168] W.R. Swartout, J.D. Moore, Explanation in second generation expert systems, in: Second Generation Expert Systems, Springer, 1993, pp. 543–585.
[169] P.E. Tetlock, R. Boettger, Accountability: a social magnifier of the dilution effect, J. Pers. Soc. Psychol. 57 (3) (1989) 388.
[170] P.E. Tetlock, J.S. Learner, R. Boettger, The dilution effect: judgemental bias, conversational convention, or a bit of both? Eur. J. Soc. Psychol. 26 (1996) 915–934.
[171] P. Thagard, Explanatory coherence, Behav. Brain Sci. 12 (03) (1989) 435–467.
[172] T. Trabasso, J. Bartolone, Story understanding and counterfactual reasoning, J. Exp. Psychol. Learn. Mem. Cogn. 29 (5) (2003) 904.
[173] A. Tversky, D. Kahneman, Extensional versus intuitive reasoning: the conjunction fallacy in probability judgment, Psychol. Rev. 90 (4) (1983) 293.
[174] K. Uttich, T. Lombrozo, Norms inform mental state ascriptions: a rational explanation for the side-effect effect, Cognition 116 (1) (2010) 87–100.
[175] J. Van Bouwel, E. Weber, Remote causes, bad explanations? J. Theory Soc. Behav. 32 (4) (2002) 437–449.
[176] B.C. Van Fraassen, The pragmatics of explanation, Am. Philos. Q. 14 (2) (1977) 143–150.
[177] 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, 2015, pp. 2469–2474.
[178] F.B. von der Osten, M. Kirley, T. Miller, The minds of many: opponent modelling in a stochastic game, in: Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI), AAAI Press, 2017, pp. 3845–3851.
[179] G.H. Von Wright, Explanation and Understanding, Cornell University Press, 1971.
[180] D. Walton, A new dialectical theory of explanation, Philos. Explor. 7 (1) (2004) 71–89.
[181] D. Walton, Examination dialogue: an argumentation framework for critically questioning an expert opinion, J. Pragmat. 38 (5) (2006) 745–777.
[182] D. Walton, Dialogical models of explanation, in: Proceedings of the International Explanation-Aware Computing (ExaCt) Workshop, 2007, pp. 1–9.
[183] D. Walton, A dialogue system specification for explanation, Synthese 182 (3) (2011) 349–374.
[184] D.N. Walton, Logical Dialogue — Games and Fallacies, University Press of America, Lanham, Maryland, 1984.
[185] J. Weiner, BLAH, a system which explains its reasoning, Artif. Intell. 15 (1–2) (1980) 19–48.
[186] D.S. Weld, G. Bansal, Intelligible Artificial Intelligence, arXiv e-prints, arXiv:1803.04263, https://arxiv.org/pdf/1803.04263.pdf.
[187] A. Wendt, On constitution and causation in international relations, Rev. Int. Stud. 24 (05) (1998) 101–118.
[188] D.A. Wilkenfeld, T. Lombrozo, Inference to the best explanation (IBE) versus explaining for the best inference (EBI), Sci. Educ. 24 (9–10) (2015) 1059–1077.
[189] J.J. Williams, T. Lombrozo, B. Rehder, The hazards of explanation: overgeneralization in the face of exceptions, J. Exp. Psychol. Gen. 142 (4) (2013) 1006.
[190] M. Winikoff, Debugging agent programs with why?: questions, in: Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems, AAMAS '17, IFAAMAS, 2017, pp. 251–259.
[191] J. Woodward, Making Things Happen: A Theory of Causal Explanation, Oxford University Press, 2005.
[192] J. Woodward, Sensitive and insensitive causation, Philos. Rev. 115 (1) (2006) 1–50.