a,∗ b b b a

Richard Dazeley , Peter Vamplew , Cameron Foale , Charlotte Young , Sunil Aryal and

a

Francisco Cruz

a

School of Information Technology, Deakin University, Locked Bag 20000, Geelong, Victoria 3220, Australia

b

School of Engineering, Information Technology and Physical Sciences, Federation University, Ballarat, Victoria 3353, Australia

ARTICLE INFO

ABSTRACT

1

2

Keywords: Over the last few years there has been rapid research growth into eXplainable Artiﬁcial Intel-

0

Explainable Artiﬁcial Intelligence (XAI) ligence (XAI) and the closely aligned Interpretable Machine Learning (IML). Drivers for this

2

Broad-XAI growth include recent legislative changes and increased investments by industry and govern-

l

Interpretable Machine Learning (IML) ments, along with increased concern from the general public. People are aﬀected by autonomous

u

Artiﬁcial General Intelligence (AGI) decisions every day and the public need to understand the decision-making process to accept the

J

Human-Computer Interaction (HCI). outcomes. However, the vast majority of the applications of XAI/IML are focused on providing

low-level ‘narrow’ explanations of how an individual decision was reached based on a particular

7

datum. While important, these explanations rarely provide insights into an agent’s: beliefs and

]

motivations; hypotheses of other (human, animal or AI) agents’ intentions; interpretation of ex-

I

ternal cultural expectations; or, processes used to generate its own explanation. Yet all of these

factors, we propose, are essential to providing the explanatory depth that people require to accept

A

.

and trust the AI’s decision-making. This paper aims to deﬁne levels of explanation and describe

s

how they can be integrated to create a human-aligned conversational explanation system. In so

c

doing, this paper will survey current approaches and discuss the integration of diﬀerent tech-

[

nologies to achieve these levels with Broad eXplainable Artiﬁcial Intelligence (Broad-XAI), and

thereby move towards high-level ‘strong’ explanations.

1

v

8

7

1

1. Introduction

3

0

Knowledge-Based Systems (KBS) researchers and designers have long understood that the ability of a system to

.

7

explain its decisions is critical to human acceptance, with approaches to providing explanations having been discussed

0

as early as Shortliﬀe and Buchanan [1] and later in a range of projects such as [2–4]. This early body of work has been

1

further applied in domains such as Bayesian Networks [5], early Neural Network systems [6] and in Recommender

2

:

systems [7, 8]. However, Machine Learning systems developed this century, such as Deep Learning, have become

v

increasingly obfuscated and non-transparent to users.

i

Recently there has been a growth in interest in eXplainable Artiﬁcial Intelligence (XAI) and Interpretable Machine

X

1

r

Learning (IML) [12, 13]. One major driver for current research has been the XAI project launched by the Defense

a

Advanced Research Projects Agency (DARPA), with twelve research programs receiving USD 75 million in funding

[14] to create a suite of explainable machine learning techniques [15]. The DARPA Project is wide ranging with aims

to develop both models and interfaces for explainability. The need for XAI has been further driven by governments

beginning to legislate requirements for autonomous systems to provide explanations of their decisions. For instance,

the European Union’s new General Data Protection Regulation [16] requires autonomous systems to be able to provide

explanations of any decisions that are based on an individual’s data. As autonomous systems increase in their level of

societal integration these legislative requirements for explanation are likely to increase. There has also been signiﬁcant

interest in XAI from futurists and innovation based companies such as AGI Innovations [17] and bons.ai [18]. Finally,

a number of conferences and workshops have been established looking at the issue from diﬀerent perspectives [12].

richard.dazeley@deakin.edu.au (R. Dazeley); p.vamplew@federation.edu.au (P. Vamplew); c.foale@federation.edu.au

(C. Foale); cm.young@federation.edu.au (C. Young); sunil.aryal@deakin.edu.au (S. Aryal); francisco.cruz@deakin.edu.au (F.

Cruz)

ORCID (s): 0000-0002-6199-9685 (R. Dazeley); 0000-0002-8687-4424 (P. Vamplew); 0000-0003-2537-0326 (C. Foale);

Miller et al. [19] suggests that AI researchers typically build explainability from their perspective rather than the

users’. There is a clear reason this is occurring. Current AI techniques are typically referred to as ‘weak’ or ‘narrow’

AI because they are designed to perform a single task. Therefore, when providing an explanation facility, the system is

only required to explain how it performed within the context of that task. For example, to explain an image classifying

Convolutional Neural Network (CNN), such as [9, 20–25], a system only really needs to identify those parts of the

image it focused on when classifying the photo or which parts of the neural network were activated. There is no need to

justify its motivations or desires as it has none. These facilities are very useful for developers to verify that networks are

focusing on the correct elements of an image to derive their decisions. Additionally, they can also provide assistance

to novice users to build some trust on individual tasks. However, with the development of increasingly integrated

intelligences, future systems will need to provide more general, broader and adaptable explanations that are aligned to

the communication needs of the explainee (section 2 will introduce a motivating example).

Improved explanation of AI decision-making and behavior will also allow for the passive education of users. Awad

et al. [26] discusses the attribution of blame when an accident occurs in mixed (human and machine) control domains.

Here it was found that people tended to blame the machine less than humans when an intervention was missed by

the secondary driver — e.g. where the AI did not take over from a human driver to avoid an accident. Awad et al.

[26] surmises that this may be due to uncertainty on the machine’s ability, while Gray et al. [27] suggests we need to

understand the sort of mind dwelling in an AI to better attribute blame and causal responsibility. Therefore, improved

explanation depth will provide better insights into the AI mind and revise people’s expectations of AI capabilities,

thereby, allowing people to make improved decisions about when they need to override control and when they can

trust the AI system to behave correctly.

In order to achieve these goals of trusted and socially acceptable systems, Miller et al. [19] emphasizes that AI

systems should be modeled on philosophical, psychological and cognitive science models of human explanation. Such

models have been researched for millennia and could be used to facilitate improved AI models for explanation [28].

Miller [28] provides three key areas of interest that should be of particular focus for good explanation: contrastive

explanation, attribution theory and explanation selection. Furthermore, Hilton [29] suggests that acceptance of an

explanation does not necessarily arise from correctness alone, but instead relies on pragmatic inﬂuences such as use-

fulness and relevance [19, 30]. Lombrozo [31] suggests that an explanation should rely on fewer causes (simple)

that cover more events (general) and maintain consistency with peoples’ prior knowledge (coherent) [32]. This prior

knowledge, Dazeley and Kang [33] suggest, builds on the situation cognition view of knowledge, suggesting that peo-

ple’s knowledge is continually reinterpreted based on their current situation and hidden contexts from within their

2

Merkwelt .

This paper argues that in order for an agent to provide acceptable and trusted explanations it must continually

determine an explainee’s contextual position through an interactive process. During this process, typically referred to

as a conversation, the agent progressively approaches the level of speciﬁcity required to lead the explainee towards the

desired level of understanding [19]. Furthermore, this paper will argue that AI cognition is structurally diﬀerent to that

of humans and that XAI systems need to go beyond cognitive interpretation alone and instead integrate technologies

to achieve human-aligned conversational explanations. This paper’s thesis will be achieved by providing deﬁnitions

of explanation levels based on the cognitive process used to generate an AI system’s decisions, and a mapping from

these to the human social process that supports the three key areas of human explanation identiﬁed by [28]. With

explanations of AI decisions mapped to human models of explanation we aim to provide researchers and developers

a theoretical basis to build better systems for trust and social acceptance. Throughout this paper we will discuss both

current work and plausible approaches to address and integrate the diﬀerent levels of explanation.

In summary, this paper makes four main contributions:

• an argument, grounded in literature, for viewing XAI through a conversational lens (section 3);

• a deﬁnition of levels of explanation with identiﬁed techniques that align with AI cognitive processes (section 4);

Until recently, the majority of AI systems existed only on computer systems and our only interaction with these

intelligences was when we chose to interact with them - such as accessing a website utilising an AI algorithm to recom-

mend purchases. Likewise, cyber-physical AI-based systems tended to be limited to highly controlled environments,

such as a manufacturing ﬂoor or sorting room. Such environments tend to be designed to exclude complexities caused

by ‘contaminants’ such as people and animals. Therefore, with no human interaction, there was little requirement for

such systems to be concerned with providing complex explanations for their behaviour. In today’s society, however,

there is already a signiﬁcant increase in the number of examples of integrated AI systems directly interacting with

people, animals and other AI agents (e.g. personal assistance, healthcare, social robotics, navigation and security). In

the coming years these systems will only increase in number and require more complex social interactions with users

and bystanders.

A common example of a relatively integrated system already in use today is the autonomous vehicle. These systems

use a range of sensors, cameras and communication systems to gather swathes of data about the environment that

they cohabit. This allows them to perceive non-autonomous environmental features like road markings, trees and

3

buildings, as well as details about other autonomous actors , such as cars, pedestrians and animals. These systems

must also maintain a memory of what has happened previously, which can be used to construct a predictive model of

other actors’ potential future behaviour. Additionally, this memory, along with a range of logically induced sub-goals

and learnt or user-set parameters, combine to provide the agent’s internal motivations, such as its goals, beliefs and

4

desires. In this paper we will refer to these internal motivations collectively as the agent’s Merkwelt (), extending

Brooks’ [37] suggestion that each autonomous AI-system has its own unique sensor suite, and therefore, will have its

own Merkwelt aﬀecting its behaviour. This extension incorporates the idea that internal motivations themselves can be

aﬀected by perceptions of the world, and therefore, will also form part of its Merkwelt. Finally, the autonomous vehicle

makes a decision about any action, or non-action, that it must take, given the current environment and its Merkwelt,

which may in turn aﬀect the environment.

While such systems work well, as soon as they are placed in mixed environments the decisions made become

increasingly complex. Importantly, these decisions become increasingly less obvious to the people with whom it is

interacting, especially when something unexpected occurs. For example, in 2017 an Uber supervised self-driving car

accidentally killed a pedestrian crossing the road [38] after apparently failing to take any avoidance measures. AI

systems such as this should be able to explain to its user, insurance companies, authorities, and the general public what

has occurred and the reasons why it made certain decisions, without having to rely on expert investigators to inspect

log ﬁles. For instance, in such a situation a car’s explanation may be:

• the system did not recognize a person pushing a bicycle;

• the system was motivated to remain in the right lane due to an approaching exit;

• the system believed that the person was giving way;

• the system believed that it had the right of way and that the person would expect the car to continue; or,

• a combination of the above.

Additionally, the system’s explanation itself and how it is generated should be veriﬁable to ensure it is not ob-

fuscating the true reasons or attempting to intentionally deceive the explainee. Of course the reason may also be a

more traditional fault such as mechanical/electrical failure, an incorrect user setting, or manufacturing fault. We are

5

not suggesting why the Uber accident speciﬁcally occurred , but rather suggesting that such systems should be able

to answer such questions themselves - as a matter of course. Furthermore, that the answer should be contextualised

3

Throughout this paper agent will refer to the AI-based system making decisions that we are primarily concerned with providing an explanation,

while actor will refer to external autonomous beings, such as people, animals and AI-agents that are separate to the agent itself.

4

von Uexküll [35] suggested that a species’ Merkwelt is determined by how it perceives the environment it inhabits, which in turn aﬀects its

behaviour and interpretation of the world. This interpretation is often extended indicating that each individual being has its own Merkwelt, such as

Environment

Static Features

• Roads

• Trees

• Buildings

• Poles

AI

Decision

• …

External Actors

• Cars

• Bicycles

Required

ℳ

• Pedestrians

Explanation

• Animals

• …

Autonomous Vehicle

Output

Figure 1: A Motivating Example providing a diagrammatic representation of an autonomous self-driving car. The envi-

ronment contains both non-autonomous features and external actors, whose behaviour may aﬀect the environment, that

need to be perceived by the agent. The agent combines the features perceived, illustrated using the collation ﬂowchart

symbol, with its internal disposition to form its Merkwelt (), which in many systems is derived from its memory of past

events and either learnt or user set parameter settings. In this example the autonomous car makes a decision in the form

of an action (or in-action), which may, or may not, aﬀect the environment. It is the reasoning behind that action and its

expected aﬀect that requires explanation.

to the person receiving the explanation. For instance, the system’s explanation to the user may be diﬀerent to that

provided to authorities and external parties. IEEE’s P7001 Transparency of Autonomous Systems working group

is developing standards around a suggested ﬁve diﬀerent types of stakeholders: users, safety certiﬁers or agencies,

accident investigators, lawyers and expert witnesses, and the wider public [40].

This example of an autonomous car and the need to explain its behaviour will be used as a motivating example

throughout this paper. Figure 1 provides a simpliﬁed abstraction of an autonomous vehicle’s interaction with its envi-

ronment, which will be used to illustrate the concepts being discussed. Similarly, the suggested possible explanations

provided above represent the range of explanation types that the various levels of explanation, introduced in section 4,

should be able to provide in a fully implemented Broad eXplainable Artiﬁcial Intelligent (Broad-XAI) approach.

3. Explanation as Conversation

Arguably, current approaches to XAI are dominated by the interpretation of the decision-making process of the AI,

rather than the provision of the reasoning about causation and expectations from those decisions and the communication

of this to a generally lay audience. This is not surprising, as interpreting a decision is a relatively achievable and obvious

ﬁrst step. These approaches allow developers and researchers to understand an algorithm’s operation in a similar way

to a good quality debugging tool conﬁrming our code operates correctly. However, such methods generally require a

level of expert knowledge to fully understand, and there is relatively limited evidence they make signiﬁcant inroads into

the generally understood aims of XAI — trust and understandability through the communication of decisions made to

diﬀerent audiences. This communication ultimately has a human on the receiving end, and therefore, we argue a full

XAI-model must integrate the human into the explanation process. To design such a model requires an understanding

Perception of Explainee’s

Understanding

Explainee’s Learning Process

Explainer’s Explanation Process

Level/Type of

Incompatible

Generalisation Required

Causal/Counterfactuals

Learning

Social

Cognitive Process

Social Process

Process

Process

Causes and

Product:

Generalised Causes

counterfactuals

Explanation

and counterfactuals

Conversation

Figure 2: Explanation as a Conversational Process , illustrating the separation of the social and cognitive processes. The

agent as a whole perceives the explainee’s understanding and requirement for an explanation, of which the social process

identiﬁes the level and type of generalisation required from the cognitive process. The explainee must incorporate the

explanation produced with its current world view and ask, explicitly or implicitly, for clariﬁcations on any incompatibilities.

ogists and cognitive scientists, what constitutes an explanation is still an active area of debate [41]. Therefore, which

aspects of explanation should be taken up by XAI researchers is also an open question. Encyclopedia Britannica [42]

describes explanation as a set of statements that makes intelligible the existence or occurrence of an object, event,

or state of aﬀairs. According to Mayes [43], “Most people, philosophers included, think of explanation in terms of

causation", but the term can also include deductive-nomological, deductive and statistical types of explanation [42].

In Psychology, explanation of human behaviour tends to be focused around a person’s social inﬂuences and internal

beliefs and desires [42, 44]. Brown [45] suggests that the roots of the term mean to ‘make plain’ and that in standard

English it most commonly means to ‘make known in detail’. Nevertheless, expatiate explanations — provision of too

much detail — should also be avoided. For instance, Miller et al. [19] point out that explanation is a component of

human conversation and that Grice’s maxims [46], for common and accepted rules of conversation, is that one should

only say as much as is necessary for the explainee to understand.

Hilton [29] presents a communication model of explanation in people as a social interaction where the explanation

must be relevant. He suggest there are two stages to an explanation: diagnosis and the explanation itself. Diagno-

sis identiﬁes causation, while the explanation is the social process to communicate the diagnosis. This aligns with

Lombrozo [47] who notes that more generally an explanation can be considered both a process and the ﬁnal product

produced by that process. While Miller [28] suggests that this process can be considered as requiring both a cogni-

tive and a social process. The cognitive process involves identifying a generalisable subset of causes, derived from

attribution, along with possible counterfactual cases. Meanwhile the social process involves an interaction between

the explainer and explainee with the aim of providing the information required for the explainee to understand the

explainer’s proposed causes of an event or outcome. Note, the explainer is only explaining proposed causes, which are

not necessarily the correct causes or even the causes the explainer believes are the correct causes [28, 47, 48]. Hence,

an explanation can intentionally be used to misinform the explainee based on the explainer’s internal goal.

Miller’s [28] concept of separating the social from the cognitive process suggests that the social process forms

what many authors refer to as the narrative-self [49, 50] or remembered-self [51, 52], while the cognitive process

relates to the experience-self [50, 52]. The latter, experience-self, refers to the fast, intuitive, unconscious process

operating only in the present situation [52]. In AI terms, this could be considered as the algorithm that is deciding

on the immediate behaviour being followed by an agent. Hence, the experience-self is the cognitive decision-making

must be a generalization and simpliﬁcation of events and their causes, it is important that a model used by an AI system

also incorporate the ability to generalise and simplify explanations that service the requirements of the explainee.

Yet, how does the explainer’s cognitive process determine the degree and type of generalization required? While

such a determination is itself a very diﬃcult open research question being pursued by numerous researchers in ﬁelds

such as culture modelling [53, 54], user modelling [55–57], emotion recognition [58–66], and human computer inter-

action [67–72], we are interested in how the determination, once made, can be used to aﬀect the explanation provided.

We propose that this question implies that the cognitive and social processes form a cycle, as shown in Figure 2. The

social process identiﬁes the level and type of generalisation required to ensure the explainee’s understanding and/or

acceptance. The social process is, thus, informed by the explainee’s reaction to its behavior or previously provided

explanation (e.g. incorrect clariﬁcation statements, questions, observed body language or tone of voice). The iden-

tiﬁed level and type of explanation required is then forwarded to the cognitive process to identify the causes and/or

counterfactuals that provide for the required level and type of explanation. Finally, the social process presents the

causes and counterfactuals to the explainee in an appropriate form.

This interaction between the explainer’s and the explainee’s social processes form the fundamental human process

of communication, which in turn informs the cognitive process’ level and types of causal and counterfactuals used.

On the explainee’s side the causes and counterfactuals are aligned to the explainee’s currently accepted understanding

of the world or more formally their objectivity illusion [73]. If the presented causes and counterfactuals are not in

conﬂict with their own then the explainee can accept them and incorporate them into their understanding of the world.

Otherwise, the explainee can identify the causes or counterfactuals that are in conﬂict with their preconceived under-

standing. The explainee’s social process can then: seek further, possibly deeper, explanations with the aim of resolving

conﬂicts; ﬁnd an alternative resolution of the conﬂicts (e.g. getting a second opinion); or, reject the explanation as

incorrect. This process of accepting or rejecting an explanation aligns with common cognitive models of how people

accept or reject arguments [74].

As previously discussed, Miller [28] suggests explanation systems should utilise existing human inspired models

of explanation to inspire greater trust between machines and people. The diﬃculty with this assumption is that AI

cognition diﬀers signiﬁcantly to human cognition [75] and extracting a human-like explanation from such algorithms

is diﬃcult. Whereas, by utilising this model of communication, where social and cognitive process are separated,

allows us to isolate the the AI’s cognitive process from how the explanation is expressed to a person. Importantly, this

also means, instead of classifying explanation systems on the ﬁnal product (the explanation), or the cognitive process

(decision-making algorithm) used; we can classify them based on the social process required to address the explainee’s

requirements. This model is put forward with the aim of aiding developers and researchers to integrate the required

systems to provide the socially required explanations.

4. Levels of Explanation

The aim of eXplainable Artiﬁcial Intelligence (XAI) is to provide explanations for decisions/conclusions made by

AI systems that people can understand and accept. Yet without a strong deﬁnition of what an explanation is in human

society, means that XAI has also been unable to provide a consistent deﬁnition for practical applications. For instance,

Gunning from DARPA’s XAI project simply describes XAI’s aim as “...ﬁnding a way to put labels on the concepts

inside a deep neural net”, as quoted by [14]. Other XAI researchers, such as Rosenthal [76] from the SEI Emerging

Technology Center of Carnegie Mellon University, state that their goal is to translate robot action, written in code,

to English. While both of these descriptions are probably simpliﬁed for lay audiences, they do represent a common

theme in the ﬁeld at the moment - that explanation primarily involves the provision of a literal translation of individual

decision points.

In ‘Explaining Decisions Made With AI’ (2020) [77], a co-badged report by the Information Commissioner’s Oﬃce

(IOC) and The Alan Turing Institute (ATI), XAI is described as consisting of both technical (extraction of information)

and non-technical (method of communication) considerations. Technical considerations are categorised as either local

Socio‐cultural (External)

Personal (Internal)

Biological (Internal

Chemical)

Figure 3: Psychological Levels of Explanation [80], illustrating the reductionist model generally used to explain human

behaviour.

may be aﬀected by the decision. Generally, we ﬁnd this binary distinction simplistic and restrictive. Hence, while

this paper focuses on the provision of individual decisions through local explanation, we recognise that these local

explanations can be supplemented with global-like explanations of sub-components that aﬀect the outcomes of the

individual decisions.

To address local explanations the IOC and ATI report [77] advocates combining and integrating explanation strate-

gies — what we are referring to as Broad-XAI. Here they describe three levels of explanation: Visualising how the

model works; Understanding the role of variables and variable interactions; Understand how the behaviours or cir-

cumstances that inﬂuence an AI-assisted decision would need to be changed to change that decision [77]. The issue

with these levels is they are highly focused on the technical aspects with the non-technical communication and human

needs being considered separately. Importantly, these levels facilitate a focus on interpretive XAI and do not consider

the social and cultural context of the AI’s decision or how to communicate this to the explainee. As discussed in the

previous section, full XAI should be considered as containing two closely interacting processes:

• social process - interacting with people, animals and/or other agents; and,

• cognitive process - identifying appropriately generalized causes and counterfactuals.

In this paper we argue that explaining individual decision points in the context of the inputs at that time, while

important in some cases, is only one part of the explanation problem. These explanations do not attempt to understand

the social context of the explainee and do not usually ﬁnd a generalisation over the causes or couterfactuals. As AI

systems increasingly become integrated into our everyday society, simply explaining a single decision point without

the historical origins of that action [79] or its social/cultural context will not carry suﬃcient meaning. We suggest that

the agent may also need to describe what it was attempting to achieve, or what other environmental factors inﬂuenced

(or did not inﬂuence) that decision.

Miller [28] suggested that before XAI can begin to produce broader explanations, we need to understand the diﬀer-

th

...

N

(Cultural)

Second‐Order…

(Social)

First‐Order Explanations

(Disposition)

Zero‐Order Explanations

(Reactive)

Figure 4: Levels of Explanation for XAI, providing a bottom-up constructivist model for explaining AI agent behaviour.

This model is adapted from Animal Cognitive Ethology’s levels of intentionality [85, 86]

.

actual explanations, they provide minimal insight into how an explanation is cognitively created, which is the problem

faced by AI researchers. Cognitive Science aims to determine the cognitive process of mind, and therefore, oﬀers more

potential. For instance, Marr’s Tri-Level Hypothesis [83] identiﬁes three Levels of Analysis: computational, algorith-

mic and implementational. These were later extended by Poggio [84] to include learning and evolution. However,

these are focused on categorising the computational approach in generating the decision rather than the generation of

the explanation.

In Psychology, any standard textbook, such as Stangor [80], identiﬁes three levels for explaining human behaviour

based on a reductionist model (see Figure 3): Social and Cultural (External Inﬂuences); Personal (Internal psycho-

6

logical) ; and, Biological (Internal chemical inﬂuences). This model suggests that explanations are provided from the

top-down, focusing ﬁrst on the social and cultural explanations and drilling down to deeper levels where required. The

advantage of this model is that it successfully divides the cognitive processes into the diﬀerent inﬂuences on a person’s

behaviour, which could work well in identifying the cognitive inﬂuences on an AI agent when making a decision. The

diﬃculty here is that AI systems do not use the same socially-focused cognitive process as people to generate their

decisions. Instead their decision-making is primarily experientially (data) driven.

Therefore, instead of this reductionist approach, we argue that a model for diﬀerent levels of explanation for AI

requires a bottom-up constructivist approach that aligns with the types of cognitive processes used by AI technology.

This goes against Miller’s [28] arguments that levels should be based on the above theories. We suggest instead that it

is more useful to have a model that aligns with AI’s cognitive process that also allows for the generation of explanations

suitable for the social process. To achieve this we propose to adapt work from the ﬁeld of Animal Cognitive Ethology

7

on Intentionality behind observed behaviour [85, 86]. Animal Cognitive Ethology has been designed to provide

explanations of animals’ behaviour based on a spectrum of intentional levels. We argue that this spectrum provides

greater alignment with current AI approaches to cognition while still providing the explanation versatility required for

social agents. Based on this approach we will deﬁne four levels of explanation plus a Meta-explanation — as illustrated

in Figure 4 and discussed extensively in the following sub sections.

Environment

Static

Static

Features

Static

Decision

Features

Entities

0

AI

Explanation

External

Actors

ℳ

th

0 (Reaction)

Agent

Output

Figure 5: Zero-order (Reactive) Explanation shown diagrammatically, where the grey oval, 0, indicates the focus area of

Reactive explanations around the AI agent’s decision based on the interpretation of the input.

4.1. Zero-order (Reactive) Explanations

As discussed earlier a signiﬁcant amount of current XAI work focuses on the interpretation of a single decision

point based on data provided to generate that decision. Sometimes referred to as local explanations [87], the aim of

these systems is to identify why the system provided a particular decision/conclusion/value/classiﬁcation based on the

data provided for that instance. In this paper, we introduce the term Zero-order explanations to formally categorize

these types of explanations, but will generally refer to them by the more descriptive name of Reactive explanations:

Zero-order (Reactive) Explanation: is an explanation of an agent’s reaction to immediately perceived

inputs.

This deﬁnition is designed to encompass all explanations that are made to describe a decision that is a reaction

to a single set of presented data or its current environmental state. In Ethology this is based on the idea that animals

may have no (zero) intentions when simply automatically reacting to a situation [86], and thus, an explanation of their

behaviour is based solely on their immediate environment. For example, a vervet monkey raises an alarm when it

perceives a danger such as a snake in the grass. If the monkey is behaving with Zero-order intentionality then it has

no motives behind raising the alarm, it simply automatically raises an alarm in the presence of danger [86]. Figure

5 illustrates the area of focus for such explanations in a typical autonomous agent. This is a generalisation of the

autonomous car example in Figure 1, where the Reactive explanation for an agent is focused on the area indicated by

the grey oval, 0, representing the process of extracting features perceived and deriving a decision.

In AI, the vast majority of current approaches are known to be purely reactive non-intentional systems, and there-

fore, any explanation provided by such a system can only ever be at a Zero-order level. For instance, a convolutional

neural network is constructed with a number of stages, each with one or more layers of neurons. The network pro-

vides a function that maps inputs to an output representing the system’s decision. By analysing the low levels you

or, ‘Why did two similar-looking cases get diﬀerent decisions, or vice versa?’.

Reactive explanations, while the lowest level, are possibly one of the most valuable. Figure 4 uses a pyramid

structure to illustrate that, while Reactive explanations are at the bottom, they are the foundation to all other expla-

nations that are built on top of this level. It is also the area currently most researched, and is commonly referred to

as Interpretable Machine Learning (IML) in the literature [89, 90]. It is imperative in any AI implementation that we

understand how a system directly interprets its immediate surroundings and inputs. These explanations allow us to

ensure that the correct inputs are being focused on and correctly processed. For instance, there are numerous examples

throughout AI development of systems that appear to have worked well, but after closer investigation, where found

to have derived the correct answer from features provided inadvertently by the engineers. An example, is the urban

legend of a tank classiﬁcation system that correctly classiﬁed whether tanks were in an image. However, upon inves-

tigation it was found the system classiﬁed images based on the direction of the shadows cast by trees on the ground.

This was reportedly caused by inadvertently photographing the tanks in the morning and the photos without tanks

being taken in the afternoon [91]. An explanation system that identiﬁes those areas of an image that were the main

focus of the AI’s attention would have quickly revealed to a human observer that it was not looking at the tanks at all.

These interpretation based explanations provide researchers and developers the ability to debug their utilisation of the

algorithm and identify how they may be able to improve performance.

There are countless examples of XAI research that can be categorized as Reactive explanation. Guidotti et al.

[89], after surveying the literature, identiﬁed that IML based explanations of black-box Machine Learning systems

fall into model, outcome or model inspection problem types. Guidotti et al. [89] suggest that people will either be

presented with a transparent model of the black-box that mimics the behaviour, such as a set of rules allowing the user

an interpretation of how the system will behave for individual cases [92]. Alternatively, the system will provide an

interpretation of a single instance’s output, such as an image, text or graph, illustrating how an individual data instance

was processed [93, 94]. Finally, the model inspection problem provides either a model or output representation of a

sub-component of the black-box, such as sensitivity to an attribute change or neural network activation levels [95, 96].

4.2. First-order (Disposition) Explanations

While Reactive explanations are the foundation of any explanation system, there is clearly a requirement for more

meaningful explanations - especially as AI systems become more advanced. As Tegmark [97] suggests, we need to

understand an agent’s goals, objectives, beliefs, emotions and memory of past events and how these aﬀect its reaction

to a set of inputs. For instance, methods such as Belief, Desire, Intention (BDI) agents and Reinforcement Learning are

already designed as goal-oriented or intentional AI systems. While these systems are still primarily reactive (e.g. given

the state of the Go board the agent makes a particular move), the way they react is based on predeﬁned or sometimes

even learnt objectives. The only way people can understand why an agent has reacted in a particular way is if the system

provides its context based on its current goal or other factors driving its behaviour. Langley et al. [98] describes this as

explainable agency and focuses on the idea that many agents are goal-directed. In particular, agency can be considered

as the capacity of an agent to act independently - making its own free choices. In terms of explanation levels we expand

on the concept of agency to be the agent’s internal disposition and deﬁne the term First-order explanations to formally

categorize these types of explanations. and refer to them descriptively with the name Disposition explanations:

First-order (Disposition) Explanation: is an explanation of an agent’s underlying internal disposi-

tion towards the environment and other actors that motivated a particular decision.

This deﬁnition builds on Reactive explanation by also considering the agent’s current internal disposition when

8 9

making its decision, such as its belief and/or desires . A Disposition explanation may, and usually will, still draw

on the foundational (reactive) explanation, but will also incorporate details of the agent’s current internal disposition

and how it inﬂuenced its reaction. In Ethology, this is based on the idea that animals’ behaviour can be based on its

8

Belief: an agent has either a predeﬁned or a learnt ‘truth’ about the environment in which it is acting. For example, an agent may have a

Environment

Static

Static

Features

Static

Decision

Features

Entities

AI

1

Explanation

External

Actors

ℳ

st

1 (Disposition)

Agent

Output

Figure 6: First-order (Disposition) Explanation shown diagrammatically, where the red oval, 1, indicates the focus area

of Disposition explanations around the AI agent’s decision, based on the interpretation of the agent’s Merkwelt, , such

as its beliefs, desires and/or memories.

current intentions that come from some form of internal belief and/or desire. For example, a vervet monkey’s belief

that there is a risk of a nearby predator or the desire for sweet fruit [86]. In these situations an explanation would

include the internal drive, such as the desire for sweet fruit (Disposition explanation), with the Reactive explanation of

perceiving a rich colour through the trees that could be fruit. In such a situation a purely Reactive explanation would

simply indicate that it saw fruit, which without a Disposition explanation would provide no meaningful justiﬁcation

for its reaction to move towards the rich colour.

Represented by the red oval, 1, in Figure 6 illustrating the area of focus, Disposition explanations allow agents

to answer questions such as: ‘why did you want to be in that lane?’; ‘why did you want to clean the room?’; or

counterfactuals like ‘why didn’t you wish to play the rook to c7?’. It is important to note the diﬀerence between the

questions asked that resulted in a Reactive versus a Disposition explanation. Reactive explanations are provided when

the question asked about how the inputs to the agent contributed to the decision. These allow us to verify its perception

and response to its inputs. A Disposition explanation, on the other hand, is provided when the question asks about

the agent’s internal motivations that caused it to react to the inputs in a particular way. Many current predictive AI

systems (such as Decision Trees, Random Forests, Convolutional Neural Networks, Naive Bayesian Classiﬁers, SVMs,

Regressions) have no inherent internal motivations other than their programmed and global need to make a prediction

based on a set of inputs. Hence, any explanation of such a system’s motivations would equate to the incorporation of

a global explanation and would remain relatively meaningless to ask such a system a question requiring a Disposition

explanation.

On the other hand, there are also a number of AI approaches that are designed to have, or can utilise, some notion of

belief and/or desire when determining a decision, such as Reinforcement Learning, Bayesian Optimization, Dynamic

Programming, BDI agents, Evolutionary algorithms, Bayesian Networks, and even search algorithms like A\*. For

instance, a Reinforcement Learning agent [99] is designed to search for an optimal solution to achieving a prescribed

ethical and safety objectives act as internal dispositions of the agent. If a user needs to understand the algorithm’s

behaviour then simply explaining the interpretation of the immediate environment is insuﬃcient and details of the

current objective, ethical consideration, or safety protocol being followed is key.

These areas of study are relatively recent and there has been very limited research in the incorporation of expla-

nation systems into these approaches. There is currently no distinct subﬁeld of XAI research that could be classed as

encompassing this level of explanation. Although recently, Anjomshoae et al. [103] suggested the term goal-driven

XAI encompassing part of this concept. The most developed Disposition explanation work is based on BDI agents

[100]. BDI agents could be considered as an extension to earlier knowledge engineering approaches to AI, in that they

10

are typically engineered solutions. Like knowledge based systems, BDI agents are grey-box systems , and therefore,

well suited to the provision of an explanation. The focus of BDI agents is the modeling an agent’s beliefs and desires

and how these change over time due to their social interactions. Therefore, they are particularly well suited to providing

Disposition explanations. For instance, Harbers et al. [105–107] developed agents that recorded behaviour-logs of their

past mental states and corresponding actions. These approaches could provide explanations for intentional behaviours

such as it opened the door ‘because it believed someone was outside’ [107]. In Reinforcement Learning and Markov

Decision Processes (MDPs) more generally, initial work in goal directed explanations can be seen in [108–110], where

causal models or predictions about the outcomes are used to explain an agent’s behaviour.

The issue with these approaches are that Disposition explanations are not exclusively derived from an agent’s

goals, objectives and beliefs, but may also be based on an agent’s emotional state, and/or its memory of past events. For

instance, many learning systems rely, not only on their current inputs, but also on a memory of past events. Disposition

explanations should be able to provide details of how these elements of their Merkwelt inﬂuence their decisions.

For instance, the family of Recurrent Neural Networks (RNNs) such as, Long Short-Term Memory (LSTMs) and

Gated Recurrent Units (GRUs), combine current inputs with a directly propagated memory of past events to determine

its outputs. Approaches have been developed to explain or provide interpretations of their decisions based on both

the input and the carried forward memories, such as Arras et al. [111] that produces heat maps during sentiment

analysis, and Bharadhwaj and Joshi [112] that explains temporal dependencies. Similarly in the recently coined ﬁeld

of Emotion-aware XAI (EXAI) an agent includes its own emotional state to explain its behaviour [113]. Human actions

are frequently expressed on the basis of one or more emotions behind an action [114]. EXAI agents ﬁrst have inbuilt

cognitive processes for formulating an emotion that is expressed through an explanation formalism [113, 115]. Memory

and emotion-aware explanations cannot adequately explain an action by themselves and, as supported by Kaptein et al.

[113], need to be combined with the agent’s beliefs, desires and goals, along with its direct interpretation of the current

state to explain its decision.

For an agent to include an explanation that incorporates both Reactive and Disposition explanations, such as EXAI

approaches [113], the algorithm being explained will generally be an amalgam of two or possibly more base level

algorithms. Currently, research into such systems may still be regarded as IML. However, we suggest that this is

misleading. Such systems need to merge multiple interpretations from diﬀerent machine learning algorithms into a

single coherent interpretation for an agent’s behaviour. Instead, we propose that such system should be regarded as

Broad eXplainable Artiﬁcial Intelligence (Broad-XAI). Broad-XAI research focuses on combining interpretations that

are extracted from multiple base algorithms. For example, Deep Reinforcement Learning [116–119] is a class of

algorithms that essentially utilise a combination of two distinct algorithms Reinforcement Learning and Deep Neural

Networks. To build a system that incorporates both Disposition and Reactive explanations, you must merge both

the interpretation of the agent’s goal from the RL transition model (Disposition explanation), with the interpretation

of the current state from Deep Neural Network performing the function approximation or feature extraction (Reactive

explanation). Currently, most explanations of Deep-RL only explain the reactive parts of the algorithm [108, 120, 121]

and would not be regarded as Broad-XAI. This is primarily because the approaches have a preset goals or objectives. As

discussed earlier approaches where the goal is learnt or derived such as Deep Multiobjective Reinforcement Learning

[122–125] oﬀer an interesting platform for Broad-XAI.

10

Agrey-box algorithm combines theoretical or engineered structure with live data [104]. A grey-box algorithm can often be directly interpreted

desire is insuﬃcient. Animal Ethology recognised that some animal behaviour may indicate an awareness of, or at

least an interpretation of, other animals’ internal beliefs and desires [86]. For instance, a vervet monkey in a zoo sees

the keeper carrying a bucket into the enclosure and concludes they intend to feed it. In Animal Ethology there is

disagreement as to whether the animal is aware of the keeper’s intentions or is simply predicting future behaviour. Re-

gardless, this and higher levels of reasoning about mental-states are recognised in humans as metacognitive processes.

This could be based on reasoned mental states as expressed in the Theory of Mind [126]; or, through imagined mental

states as expressed by the concept of mentalisation [127] — colloquially often referred to as empathy, emotional un-

derstanding, attribution, mind-mindedness and self-awareness. Mentalisation can be considered as the ability to see

ourselves as others see us, and others as they see themselves [128]; or, more formally, it is the recursive understanding

of mental states [129]. Both this and the next level of explanation in this paper are based on deeper recursive levels of

mentalisation. Current evidence suggests that mentalisation may be unique to humans [130, 131] and can be regarded

as core to social cognition in humans and is thought to develop in children around the age of 4 or 5 years [129].

As AI agents become increasingly integrated within our society, it is clear that in order to function and be accepted,

they will require the ability to model other actors’ (people, animals and agents) behaviour [132–135]. For example,

in order for the Uber self-driving car, discussed earlier, to function the system requires a model of behaviour for other

actors in the environment — such as the pedestrian with a bicycle. However, if an agent simply models behaviour

then when providing an explanation it can only discuss the expected behaviour, but not provide any reasoning about

why it expected that behaviour. In order for an agent to describe another actor’s intentions, it will also need the model

to predict the actor’s internal state and likely memory. This model provides a belief of other actors’ mental states

and allows for the prediction of both their behaviour and their likely motivations for that behaviour. Agents that are

provided with or learn such a model will also require the ability to explain decisions made based on this model. In this

paper, we deﬁne a Second-order explanation to formally categorize these types of explanations, and will generally be

referred to by the more descriptive name of Social explanations:

Second-order (Social) Explanation: is an explanation of a decision based on an awareness or belief

of its own or other actors’ mental states.

This deﬁnition builds on Disposition explanation by also considering the agent’s model of other actors’ internal

mental states. This idea is represented visually in Figure 7, where the three purple ovals indicate the areas of interest, 2 ,

a

2 and 2 . In order to provide this level of explanation, an agent must ﬁrst assume that other actors in the environment

b c

11

have their own Merkwelt , 2 , that guides their behaviour. The agent must also be engineered, or able to derive/learn

a

through observation, how changes to the environment by the agent, 2 , and other actors, 2 , may inﬂuences each actor’s

b c

internal state. It is this predictive model, which in turn feeds the agent’s decision-making, that must be explained by a

Social explanation. This model of how each separate actor is expected to behave is essentially a model of that external

actor’s Merkwelt, or at least how that manifests in its outward behaviour, and will be referred to in this paper as the

Actor’s Model (). It should also be noted that this can also include a model interpreting the agent’s own mental state

as it believes others would see it.

This level is referred to as Social explanation as it aligns with basic human reasoning in social interactions, whereas

the lower levels of reasoning simply involved an agent responding to its environment and achieving deﬁned tasks. Peo-

ple’s only interaction with such systems is to maintain the assigned tasks and/or take care to avoid the agent’s operating

environment. These systems work well in domains such as: autonomous packing facilities, where there are no peo-

ple involved for the agent to interact; or, domestic vacuum cleaners, where people can reasonably avoid the devices’

wanderings. In these contexts of low-level reasoning, Reactive and Disposition explanations will suﬃce. However,

higher order reasoning is required for devices such as robot carers, autonomous self-driving cars, autonomous per-

sonal organisers, etc. Agents performing these higher levels of decision-making must be able to explain the reasoning

involved in making those decisions. For a user to understand that reasoning, any explanation should also provide the

AI’s reasoning about other actors’ dispositions.

2

b

Environment

Static

Static

Features

Static

Decision

Features

Entities

AI

Explanation

External

Actors

2

c

ℳ

nd

2 (Social)

Agent

Merkwelt

Merkwelt

Output

𝒜

2

a

Figure 7: Second-order (Social) Explanation shown diagrammatically, where the purple ovals, 2 and 2 show that the

b c

agent’s and other actors’ behaviour aﬀects the environment, which in turn aﬀects the Merkwelt of the agent itself and the

other Actors’ Models (). The agent models the other actors’ potential internal state and memory, represented by the

oval 2 , to predict their future behaviour.

a

Social explanations allows an agent to answer questions like ‘why did you slow down when the pedestrian ap-

proached?’; open questions that do not imply a level, such as ‘why did you move my meeting time?; or, counterfactuals

like ‘how would you have played if I had raised by a larger amount?’. The answers provided by such a system would go

beyond its own intentions and explain how other actors’ aﬀected its behaviour. For instance, for the last three questions,

it may provide explanations like: ‘I slowed down because the pedestrian approaching the intersection appeared intent

on crossing’; ‘I changed your meeting time because I believed you would not wish to get up early after watching the

late football game the night before’; or, ‘Had you raised by more then I would have thought you were bluﬃng’.

There are a number of factors in developing systems that can cognitively build a model of another actor. For in-

stance, this may involve: the attribution of disposition or situational contexts to an actor’s behaviour; cultural modelling

of an actor; reﬂection on one’s own beliefs and biases; emotion modelling of actors including the agent itself; and, the

causal attribution of an actor’s actions, or lack there of, to events or non-events. Furthermore, this is highly dependent

on the type actor being modelled, for instance a person compared to an animal or another autonomous agent require

diﬀerent modelling approaches. Each situation involves developing a model for each actor’s internal state or memory,

including the agent’s own as others may believe it to be. This modeling of external actors is an active research ﬁeld

with substantial work already caried out. For instance, culture modelling [53, 54], user modelling [55–57], emotion

recognition [58–66], action recognition [136–142], pedestrian prediction [143–148] and human intention modelling

[69–72] are just a few areas that could be considered as modelling external actors.

There are a number of examples of applying IML based techniques to these Actor Models. One common application

is to use XAI approaches to interpret a video stream to explain what had occurred or to provide captions [149–153],

or a narrative [154]. Alternatively, researchers use Local Interpretable Model-Agnostic Explanations (LIME) [155] to

generate explanations of a range of Actors’ Models, such as [156, 157]. As mentioned, all of these XAI systems are

However, Social explanation is not particularly interested in these methods per se, but rather, in providing an expla-

nation for how these methods’ conclusions contributed to the agent’s behaviour. A system providing an explanation of

how it interpreted an external Actor’s Model is essentially still only providing a Reactive explanation of that predictive

model. This is because these approaches are applied domains for general machine learning and artiﬁcial intelligent

algorithms. An XAI system providing a Social explanation may include the interpretation of the actor’s model, but

more importantly will describe how that outcome aﬀected its ﬁnal decision. Viewing Social explanation in this way is

aimed at simplifying how such a system can be.

Despite the prevalence of actor modelling and augmentation there is relatively little work taking the next step

of providing an explanation of how the augmented model has aﬀected the agent’s behaviour. In our discussion on

Disposition explanation we discussed the recent introduction of Emotion-aware XAI (EXAI) [113]. This paper not

only discussed explaining the agent’s internal emotions but also introduced the idea of explaining external actor’s

emotions and how they aﬀected the agent’s decision-making. More broadly than emotion augmentation, there has

been work in mental state abduction [168–170] of BDI agents, where there are examples of explanations describing

how these deduced mental states have changed an agent’s behaviour. BDI research in this area is primarily focused on

identifying the mental state of another BDI agent, for which it already knows the possible mental states the agent can

have, rather than modelling the mental state of less deterministic external agents, such as humans.

th

4.4. N

-order (Cultural) Explanations

The social level, section 4.3, in human terms represents a relatively simple degree of interaction. People automati-

cally expect others to determine what they are doing and adjust their behaviour accordingly. The issue is that this level

of reasoning makes no assumptions of the need to provide some degree of quid pro quo. That is, if one person gets

out of another’s way, is there a reciprocal responsibility? In a strongly hierarchical society — perhaps not. In many

human-agent interactions some may assume that we want such hierarchies — where agents always adjust behaviour

to stay out of people’s way. However, this could result in agents being unable to complete required tasks as they must

always avoid people. As most agent tasks are being carried out as proxies for a person, that person may not wish for

their agent to always be subservient and defer to humans. In our motivating example, imagine that the autonomous car

always gives way to all other cars in every situation. Such a car may never go anywhere and would quickly become

useless to the human passengers. Furthermore, it will not behave like human drivers and, thereby, create confusion

and risk as it does not meet other drivers’ behavioural expectations.

In a complex society people do not just develop a model of other’s and then adjust their behaviour to avoid them.

Instead, they also derive a set of expectations of how those other actors’ believe they will behave to avoid them. For

instance, if two people walking through a shopping center are heading directly towards each other, person A may

decide that they should side step to the left slightly while also having an expectation that person B will do similarly

creating enough distance between their paths to avoid the impending collision. This level of reasoning is equated to

third-order intentionality [171] because person A, not only has a model of what person B will do, but recognises that

person B will expect person A to do likewise — and vice versa. This represents an ever increasing recursive level of

mentalisation [128, 129] indicating an understanding of a set of cultural rules about behaviour. For instance, the above

example was based on the cultural assumptions used in Australia — if person A is visiting the United States of America,

where people step to the right instead, taking a step to the left is likely to create confusion and possibly be the cause of a

collision. This represents a shared cultural and social awareness. With people that know each other personally this same

awareness may go beyond general cultural norms and consider a deeper and personal understanding of expectations.

Current First-order intentional agents like vacuum robots are built with the assumption that people will entirely

defer to them and get out of their way. People tolerate this due to our low level of expectation of the device’s intelligence

and because avoiding one such device is simple. However, with more devices involved in more complex interactions

greater expectations will be placed on the agent to not only avoid us, but to culturally integrate with us. Therefore,

agents will be required to, not only be provided with or learn a model of how other actors will behave, but to also

determine what expectations those actors may have of how the agent should behave. Ultimately, this is recursive, with

Environment

Static

Static

Features

Static

Decision

Features

Entities

AI

N

b

Explanation

External

Actors

ℳ

th

N

(Cultural)

N

a

Agent

Merkwelt

Merkwelt

Output

𝒜

th

Figure 8: N -order (Cultural) Explanation shown diagrammatically, where the green oval, N indicates that we need to

a

model how the actors’ potential internal state and memory are aﬀected by changes in the environment, while, N indicates

b

the need to explain how the changing environment aﬀects the agent’s Merkwelt.

descriptive name of Cultural explanations:

th

N -order (Cultural) Explanation: is an explanation of a decision made by the agent based on what

it has determined is expected of it culturally, separate from its primary objective, by other actors.

This deﬁnition builds on Social explanations by also considering the agent’s model of other actors’ internal model

of the agent’s behaviour, which has been built or learnt through observation. In other words, a Cultural explanation

details how the agent’s behaviour is modiﬁed due to what it understand is expected of it by other agents. This idea

is represented visually in Figure 8, where the two green ovals, N and N indicate the areas of interest. Here it can

a b

be seen that we are interested in modelling both how the external actors’ models are altered, N , as well as how the

a

agent’s own Merkwelt, N , is aﬀected by changes in the environment. Such an explanation requires the agent, to not

b

only model the other actors’ potential internal state and memory, but also how the actor’s model may be aﬀected by

changes in the environment.

This level is also referred to as Cultural explanation as it aligns with basic human reasoning in cultural interactions

and is aligned to the philosophical theory of Sociality: incorporating the idea of a collective intentionality [172].

This implies that any expectations determined for one Actor’s Model or situation may potentially be transferable or

shareable with other Actor’s Models or situations. The types of questions answered by this level are the same as those

in Social explanations, except now the provided explanation incorporates its cultural understanding of others’ expected

behaviour. For instance, a cultural understanding of one road user in France can, for the most part, be transferable to

all road users in France, but only partially transferred to road users in China. Cultural explanations would, therefore,

be able to provide explanations such as: ‘I waited a moment to hold the door open for the person behind me, as

social obligations indicate that they would expect this behaviour’; or, counterfactuals like ‘I didn’t slow down for the

pedestrian because road laws state I have right of way, and therefore, I assumed they expected me to continue’. These

actor expects the agent to do and how they are likely to respond to the agent’s behaviour. This form of behaviour

modelling has been widely studied. For instance, Normative Agents research has extensively studied multi-agent

ﬁelds like BDI agents [173–176]. These approaches focus on constructing multi-agent societies where social norms

(of the multi-agent society) can be applied to the expectations of the agent. However, like other BDI work these

systems typically avoid operating in dynamic human environments. Game Theory [177], which models the strategic

interaction between rational decision-makers, assumes that other actors will behave optimally to achieve their goal.

Game Theory has been extensively studied and applied to numerous real-world domains including modeling social

and cultural interactions [178, 179]. One common use is in modelling opponents in games. For example, Alpha Go

[180], which defeated the World Champion Lee Sedol in 2016, utilised a Monte Carlo tree search (MCTS) [181] to

predict responses by its opposition on various possible actions it could make. Furthermore, it can build a probability

to do this over a sequence of actions and not simply model the next behaviour. Modelling expectations has been more

directly modelled in social action research, which models when, why and how external demands on an agent aﬀect

its goals or actions [182–184]. Other learning systems, such as Reinforcement Learning are sometimes designed to

incorporate social and cultural awareness of the environment into their action selection, for instance, when navigating

through crowded domains [185–190].

Nevertheless, these areas have not currently attracted signiﬁcant research interest in the development of explanation.

For instance, David Silver, DeepMind Researcher, reportedly had no insight into why AlphaGo made such a creative

play as move 37 in Game 2, which surprised both commentators and Lee Sedol, until he had investigated the actual

calculations made by the program [191]. For these systems to go the next step and be used by everyday non-expert

users, that are not able to inspect an agent’s value-function, agents must be able to provide explanations at this level.

Recent research by Kampik et al. [192] into explaining sympathetic actions that incorporate utility for socially beneﬁcial

behaviour at the detriment of personal gain and using this to explain behaviour is an exciting example of culturally-

aware explanation. Other examples of explanation systems based on cultural expectations do not align to a unique ﬁeld

of research, but are occasionally published under more general terms such as understandability [193], transparency

[194], predictability [195].

Interestingly, when an agent begins to base its decisions on this idea of quid pro quo, it not only means there are

expectations on the agent but also on the other actors. Like all socially-aware actors they will also have expectations to

follow. Understanding and modeling this means an agent can go beyond simply predicting how an actor’s model will

change based on the agent’s behaviour, and instead, model how it can actively change the actor’s behaviour. Research in

Computers As Persuasive Technologies (or Captology) investigates approaches to persuade external actors. Captology

is already used in a number of application domains, such as commerce, safety, environment, productivity, disease

management, health care, activism and personal management [196–200]. Captology implies a goal or intention to

alter and change other actors’ internal beliefs, motivations and behaviours and can be considered as Fourth-order

intentionality [171]. An explanation for such a system would amount to a description of how an agent’s action was

based on the goal/intention to change an actor’s beliefs, motivations and/or behaviours in a particular way. In this way

the explanation could be generated using a Disposition explanation approach. We collectively have wrapped these

terms up into Cultural explanation in this paper as levels of intentionality, in theory recursively continue. To our

knowledge there have been no attempts to provide explanations for such systems.

4.5. Meta (Reﬂective) Explanations

One of the major drivers in the development of XAI is to provide understanding to people so they can trust the

decisions made. However, a major issue in XAI research, which is generally not discussed, is if we cannot trust the

agent’s original decision, how can we trust the agent’s explanation of that decision? The majority of research in XAI has

been researcher/developer-centric [201] — providing an explanation that focuses on interpreting the decision. As these

are direct interpretation methods they tend to be more readily acceptable to an expert, but do not generally provide the

level of understandability required by a non-expert. Human-centered interaction methods instead approach the problem

from the human perspective by focusing on the user’s acceptance and understanding [12, 201, 202]. These approaches

3. Simpliﬁed/generalised explanations - inadvertent deception through omission.

Discussed in section 3 the narrating-self, the conscious part of the mind, creates explanations that support the

internal goals of the individual. Therefore, the aim of the explanation could be considered in terms of maximising the

agent’s utility. This aligns with a human-centered approach to explanation, where the aim is to measure the utility of

an explanation on the human’s perceived level of acceptance or understanding and to use this to reﬁne the explanations

provided [201–204]. In fact, the primary objective when providing an explanation could be considered the requirement

to satisfy the explainee’s need to understand and accept a decision — not to accurately reﬂect the actual decision-making

process. This requirement to provide an understanding to the user is potentially in direct conﬂict with providing a

‘truthful’ and accurate explanation. An agent that learns how to explain a concept may learn to use an explanation

that returns the highest utility, rather than the most accurate. For example, Emotion-aware XAI was introduced as

an approach to both explain an agent’s internal disposition (section 4.2) as well as, interpreting an external actor’s

disposition (section 4.3). One approach used to develop such agents is to use emotion-driven or emotion augmentation

learning methods [158, 162, 205–209]. These methods often use the emotion of the agent or the perceived emotion of

an actor to control its behaviour, such as an agent’s reward being aligned to pleasing its user. If this approach is used

when providing explanations, then an agent will learn an explanation that satisﬁes the user — not the truth.

Wright et al. [210] recently introduced the idea of rebellious and deceptive explanations. These are explanations that

are generated, not just to explain a decision, but to intentionally persuade a person and, thereby, change their internal

beliefs and desires. Such explanations are not frequently discussed in the literature, but represent a signiﬁcant safety risk

to one of the primary objectives of XAI — developing trust. Deception can be in the obvious form of lying [211–213],

but as Sakama [214] suggests it can also be a result of bluﬃng or truth-telling (agent makes a truthful explanation that

deceives the listener) [214, 215]. Intentionally deceiving agents for instance can be used during negotiations, such as

where an agent representing a company may need to deceive a customer to advantage the company. Current approaches

[212, 216, 217] focus on logic formalisms for deception, but future work in explaining black-box Machine Learning

algorithms are likely to also encounter this concept in future research. Sakama et al. [218] suggest four Quantitative

and Qualitative Maxims for Dishonesty that build on from Grice’s [46] maxims for common and accepted rules of

conversation. Sakama et al.’s [210, 218] maxims suggest an agent should:

1. Lie, Bullshit (BS), or Withhold Information (WI) as little as possible to achieve its objective

2. Never lie if it can achieve its objective by BS.

3. Never lie, nor BS if it can achieve its objective by WI.

4. Never lie, BS, nor WI if it can achieve its objective with a half truth.

Deception does not necessarily need to be designed - it could also be a learnt trait of an agent. The last paragraph

of section 4.4 discussed the research ﬁeld of Captology, which can be considered a Fourth-order intentional systems.

These systems are designed to persuade and alter an actor’s beliefs, motivations and/or behaviours. This ﬁeld of

research was introduced as an example of the recursively higher-levels that a Cultural explanation system ultimately

may need to explain. However, it also introduces a troubling concept: what if the agent’s method of persuasion is

through the explanation itself?

A deceptive explanation does not only result from an agent having the intention to deceive, or to maximise a utility

for human acceptance and understanding, but can also be the result of a poorly formed explanation. For instance,

Grice’s [46] maxims for common and accepted rules of conversation, suggests an explanation should be as brief and

as general as possible. One cause of miscommunication, and potentially unintentionally deceptive explanations, is

through the omission of some information to maintain brevity, causing a misunderstanding on the receiver’s end. The

process of generalisation can also result in the explanation misrepresenting certain facts about a situation.

In this paper we suggest that human trust requires more than just the explanation, but also requires an understanding

of the process used in formulating the explanation. That is, how was the explanation generated or selected, as opposed

to other explanations, and importantly, how the explanation provided is related to the decision made. We suggest that a

Environment

Static

Static

Features

Static

Decision

Features

Entities

AI

Explanation

External

Actors

ℳ

Meta

Agent

Merkwelt

Merkwelt

Output

𝒜

Figure 9: Meta (Reﬂective) Explanation shown diagrammatically, where the yellow circle, Meta, encompasses the ex-

planation levels indicating that a Meta-explanation is concerned with inferring how each of the explanation levels were

mapped.

Meta(Reﬂective) Explanation: is an explanation detailing the process and factors that were used

to generate, infer or select an explanation.

Meta-explanation is not a commonly used term in the XAI literature. Generally, it has been used to refer to an

explanation of Meta-knowledge [219, 220]. That is, rather than explaining why a decision is the correct decision (the

explanation) it explains the problem solving process in solving it. In systems based on logical inferences this often

referred to an explanation based on a rule-trace [220]. More broadly than just AI, the literature has also separated ex-

planation into two levels: Object-level explanation and Meta-explanation. The Object-level provides an explanation of

particular decisions or arguments, whereas a Meta-explanation references the scenario structure behind the explanation

or argument and may include historical decisions or justiﬁcations [221]. Up until now, we have suggested four levels

of explanation that equate to the Object-level explanation used in the wider literature. A Meta-explanation equates

to the wider literature use of the term except that we propose to broaden these ideas speciﬁcally for use in XAI and

suggest that a Meta-explanation, not only describes the problem solving process or the scenario structure, but also the

process used to select or generate the explanation of the original decision. This is illustrated diagrammatically in Fig-

ure 9, where all explanations previously discussed represent the internal processes requiring explanation. Therefore, a

Meta-explanation, the yellow circle, Meta, is drawn from all other forms of explanation.

Meta-explanation was deliberately illustrated in Figure 4 as running across and alongside each of the explanation

levels. The aim was to indicate that a Meta-explanation can form a symbiotic relationship with each of the levels.

Essentially, we are suggesting that an explanation may be just a level explanation or could also contain part or entirely

Meta-explanation details. For instance, a Reactive explanation of an agent with Zero-intentionality is also considered

a Meta-explanation. This is because a Reactive explanation is a direct interpretation of the algorithms processing of its

reaction to an input. The objective of the explanation is to illustrate the problem solving process. This is a traditional

form of Meta-explanation that is often called Interpretable Machine Learning (IML) [89, 90]. For example, asking an

This is due to Grice’s [46] suggestion to simplify an explanation and, thus, representing a Reactive explanation without

including the Meta-explanation. For example, in the self-driving car example, in section 2, the ﬁrst explanation suggests

that the agent did not see a person pushing a bicycle. This is a Reactive explanation without a Meta-explanation. Should

the explainee require further information about how the car reached the conclusion that there was no person with a

bicycle they would need to request a deeper explanation such as a Meta-explanation of the process used by the agent.

This separation of level-explanation and Meta-explanation is clearer in the higher-levels as these are typically al-

ready providing Broad-XAI explanations, and therefore, asking how it determined an explanation becomes more mean-

ingful. For a fully Broad-XAI explanation that goes across all levels of the hierarchy providing a Meta-explanation

becomes an important component of drilling down to validate the process used by the agent in determining the expla-

nation given. For instance, if the self-driving car gave the third example explanation that it thought the person was

giving way then the Meta-explanation would need to provide an explanation of the agent’s process used to determine

the person was giving way.

The exact nature of a Meta-explanation is directly determined by the method used to generate the original expla-

nation. For instance, if an explanation is determined from a set of predetermined options then the Meta-explanation

will provide details of the classiﬁcation process used. Then again, if the explanation uses probabilities based on past

situations then the Meta-explanation would need to detail or trace the approaches used to determine the probabili-

ties calculated. Essentially, a Meta-explanation could be considered as the backing to the warrant (explanation) of

Toulman’s argumentation model [74].

Given this deﬁnition of a Meta-explanation as being purely an interpretation of the process used in forming the

explanation then the Meta-explanation is also disconnected from the agent’s learning, memory and reward structures.

This means the resulting explanation can be seen as a truthful (accurate) reﬂection of the agent’s behaviour. This is an

important diﬀerentiation as it separates a Meta-explanation from the ﬁrst issue raised by utility-driven XAI. A regular

explanation is aiming to satisfy the explainee’s need to understand the reasons for the decision and not the processing

method used. If meeting this need simpliﬁes, generalises or omits information about the agent’s true objectives then

the additional provision of a Meta-explanation, while potentially less understandable, provides an honest and accurate

reﬂection of the decisions and explanations provided (Issue 3). To this end we actually see such a Meta-explanation as

outside the normal conversational component of an explanation and as a method of providing a deeper understanding

to the explainee if required.

Unfortunately, we believe it is unlikely that a Meta-explanation will be provided by certain systems. Intentionally

deceptive agents (issue 2) created by organisations will frequently require the exact decision-making process and its

associated explanation generation to be hidden from clients. Forcing companies to provide a Meta-explanation is

an issue to be taken up with governments — potentially the same ones that will be unwilling to provide such details

themselves. For instance, the new General Data Protection Regulation [16] aims to ensure such processing components

are made available to those aﬀected by decisions made by automated systems. There is a risk that companies could

view the provision of a regular explanation as suﬃcient. Instead, we believe that a Meta-explanation of these decisions

are also required.

5. AConceptual Model for Providing Explanations Through Conversation

The aim of this paper has been to present a conceptual approach to bringing together diﬀerent AI and XAI compo-

nents to guide and inspire future research towards the development of Broad-XAI based conversational explanations.

In so doing we have reviewed a number of disparate research ﬁelds and through a structured set of levels of explana-

tion illustrated how these separate ﬁelds can provide a uniﬁed approach to XAI. In particular, this paper has rejected

current interpretable machine learning (IML), or debugging-based, approaches as a solution in-and-of-themselves for

XAI. We have argued that these approaches are primarily focused on providing Zero-order (Reactive) explanations.

Higher levels are identiﬁed that allow for more socially and culturally-aware explanations to provide a human-aligned

explanation.

Explainee’s

Understanding

Explainee’s

Explainer’s

Social Process

Social Process

Socio‐cultural

th

N

(Cultural)

Personal

Second (Social)

Biological

First‐Order (Disposition)

Zero‐Order (Reactive)

Product:

Explanation

Conversation

Figure 10: Conversational Interaction between the levels of Explainable AI and human psychological levels (builds on top

of Figure 2). When the explainee requests an explanation (black arrows) at a Socio-cultural level the agent responds from

th

a Second or N -order level of explanation. Similarly, when requesting an explanation at the Personal or Biological level

the agent will respond from a First or Zero-order level of explanation respectively. In each case, the explainee can also

request a Meta-explanation (grey arrows) on how a prior explanation was generated.

application requirements. In so doing we will brush over the minutia of such an implementation, as these should be the

focus of future research. The following sub-section will reprise the earlier discussion on explanation as conversation

and present a process for performing an explanation through an interactive conversational process. Subsequently, the

following sub-section will provide an insight into the AI-model required to provide the cognitive processes required

to for Broad-XAI and the identiﬁed levels of explanation. The ﬁnal sub-section will provide a brief review of current

approaches to developing Broad-XAI approaches and discuss how they relate to this paper’s proposed approach.

5.1. A Broad-XAI Conversational Process

In section 3, we discussed how an explanation was a process consisting of both a cognitive and social process,

where the social process involves an interaction between the explainer and the explainee. The aim of this interaction

is to convey the explainer’s proposed causes of an event or outcome such that the explainee understands and accepts

the explanation. This interaction forms the fundamental human process of communication, which in turn informs the

explainee’s cognitive process. On the explainee’s side of this communication the identiﬁed causes and counterfactuals

are aligned to the explainee’s currently accepted understanding of the world. The explainee can then accept or reject

the provided explanation. Alternatively, they may require the agent to provide backing claims [222] by requesting

additional, more specialised or detailed explanation to resolve any identiﬁed internal conﬂicts.

Furthermore, it was identiﬁed that the psychological models of explanation in humans uses a reductionist model

(see Figure 3) suggesting explanations are provided and received using a top-down approach. In contrast, we have

argued that AI systems do not operate in the same socially-focused cognitive process and instead utilise a data-driven

constructivist approach. We, therefore, presented a set of levels that builds explanations from the bottom-up. Ulti-

Assumptions

Explainee’s

Quiescence

Interpret

Explanation

Question

Initial

Acceptance /

Question

Rejection

Clarifying Question

Figure 11: The Conversational Process iterates through three stages until the explainee reaches quiescence - A utility

measure of understanding and acceptance. The cycle terminates when the explainee accepts or rejects the explanation

and moves on to either a new line of questioning or no longer requires an explanation.

an explanation is to identify the appropriate level from the approach used by the explainee when phrasing the request

for an explanation. Figure 10 provides an illustration of how an agent’s levels of explanation aligns with that of a

person’s psychologically-based levels by embedding these models within the conversation process shown earlier in

Figure 2. For instance, let us assume the user asks ‘what did the vehicle perceive just prior to the collision’. Clearly

such a request is for a Reactive explanation of the agent’s perception to determine the reason for the reaction. Generally

speaking, we expect it would be unlikely that the user will phrase a request for a speciﬁc level. Rather, a more likely

assumption would be that the explainee would ask ‘why didn’t the car give way to the pedestrian pushing a bicycle’?

Such a question is not suﬃciently aligned to any individual level of explanation. In fact, any of the four responses

indicated in section 3 would provide a reasonable explanation — although alone they are unlikely to fully satisfy the

explainee.

One approach would be to wrap-up all the explanations from each level into a single explanation. For example, a

combined explanation may look something like:

‘The system perceived the pedestrian with a bicycle approaching, however, was motivated to remain

in the right lane due to an approaching exit, believing the person would give way to the vehicle and,

as the car had the right of way, that the person would expect the car to continue’.

However, this would break Lombrozo [31] and Thagard’s [32] suggestions that an explanation should rely on as few

causes (simple) as possible that cover more events (general) and maintain consistency with peoples’ prior knowledge

(coherent). The combined explanation makes no determination of the explainee’s prior knowledge; is very speciﬁc

to the exact situation; and, complicates the explanation with many causes that may be irrelevant. Therefore, such an

explanation, we believe, would not be suitable.

Interactionism in sociology studies the social process: between two people; between a person and an animal;

increasingly, between a person and an AI-driven agent [223–225]; and, even between animals and an AI-driven agent

[226]. Interactionist theories suggest that multiple levels of explanation are necessary to explain a particular behaviour

and that often the interplay between these levels is also required [227]. This suggests that an explanation should be

constructed from several levels. We envision that this could be done in an incremental approach such as through an

interactive conversation. An agent initially would provide an explanation at the highest level of the agent’s intentionality

that answers the question and drills down through the levels to answer subsequent queries until the explainee accepts

own world view. However, the conversational explanation process can also terminate, even when quiescence is not

reached, by the explainee rejecting an explanation entirely. Such a situation occurs when the explanation does not

align with their personal world view. Similarly, the explainee may end the process because of frustration or to move

on to a diﬀerent line of questioning. The measurement of this type of utility has been considered through examination

dialogues [228–230]. This process equates to Fourth-order level of intentionality where the agent is actively trying to

move the explainee’s understanding towards a level of understanding and acceptance.

The ﬁrst stage is to correctly interpret the question (either explicit or implicit) conveyed to the agent. On the surface

this appears to be simply an interpretation of their words or body language. However, to provide a conversational

explanation the agent needs to understand the underlying nature of the explainee asking the question. Therefore, we

suggest the agent requires a model of the explainee’s current understanding of the world in the form of an Actor’s Model.

Such a model allows the agent to correctly interpret the true nature of their request for an explanation and to provide

a more accurate explanation. This stage illustrates a relationship with Second-order intentionality, where the agent’s

choice of behaviour, the explanation, is determined by a model of the explainee. User modelling in Human Computer

Interaction (HCI) [67–72] approaches can be used to model the explainee allowing more meaningful interpretations

of the questions. This is evident in recent work in the emerging ﬁeld of Personalised Explanations [231]. At the

conclusion of this stage the agent should have a, possibly re-framed, question that is then addressed in the next stage.

The second stage aims to identify and conﬁrm or clarify any assumptions that exist in the question that may be

incorrect. These are important because if the explainee has an assumption about the agent’s reasoning that is incorrect

then identifying that assumption and pointing out in what way it is wrong is likely to address their primary concern.

For example, when asking ‘why didn’t the car give way to the pedestrian pushing a bicycle?’ there is an assumption

that there was a pedestrian pushing a bicycle and that the car had detected this. If this was not detected then the agent

should jump immediately to a Reactive explanation and identify what it did, or did not, perceive. Essentially this

involves rephrasing the question to be speciﬁc to the lower level of intentionality ready for the third stage. If there are

no assumptions in the question, or the assumptions are correct, then the question can be passed to the next stage.

The ﬁnal stage is to produce an explanation based on the highest relevant level to the question asked involved

in the original reasoning process that has not yet been explained. Therefore, assuming the agent’s original reasoning

process involved all levels of intentionality and that the question is open to a response from any level then the agent will

respond from the highest level of explanation. For instance, the car ‘believed that it had the right of way and that the

person would expect the car to continue’. The elegance of this explanation is that, in a single generalised explanation,

it has implicitly also conﬁrmed that the assumption that there was a pedestrian pushing a bicycle was in fact perceived

by the agent and the car had modelled both its responsibility and determined a prediction of the pedestrians expected

intentionality.

Once this explanation has been oﬀered the agent measures the explainee’s quiescence. In the event that the ex-

plainee has not reached this state yet then we repeat the cycle with a new question. We suggest there are three forms

for this subsequent question:

• They ask a question that drills down further. For instance, ‘did the pedestrian show any sign they intended to

give way?’. Such a question will generally require a lower level of explanation on the next cycle.

• Their body language or facial expression indicates some level of confusion resulting from the explanation given.

In which case the agent can oﬀer an explanation from a lower level to provide greater clarity.

• They ask an alternative branch of questioning. For example, ‘Why were you in the right lane?’ A totally new

branch indicates they have exhausted the prior line of inquiry and now wish to start a new explanation cycle.

In this way, the agent can provide general explanations initially and progress towards more speciﬁc explanations

until the explainee is satisﬁed. This approach addresses Miller’s [19, 28] three key areas of interest in a quality expla-

nation system: contrastive explanation, attribution theory and explanation selection. For instance, this approach builds

Merkwelt

Model

An explicit or implicit model of the

agent’s own internal motivations,

such as its goals, beliefs and desires.

N

Explain how the Merkwelt affects the

1

agent’s behaviour/actors/ environment

using goal & emotion-aware

approaches.

Behaviour

Perception

Model

N

Process to extract features that

An engineered or learnt model

represent the environment. For

mapping perception, actor and

instance ANNs, Fuzzy sets,

M

0

M

0

Environment Behaviour

Merkwelt models to behaviours.

probabilities etc.

Explain the agent’s choice of behaviour

Explain using IML approaches such

Explain the expected effect of

N

using introspective, contrastive and

as feature summaries, saliency

the agent’s behaviour on the

counterfactuals the agent’s

maps, statistics, etc.

environment

action

Actor’s Model

A model for each actor in the en-

2

vironment used to predict each actor’s

N

potential future behaviour.

Explain how the agent’s belief about each

actor’s predicted behaviour affected

the agent’s Merkwelt, Behaviour

and the environment using

Social & Cultural

approaches.

M

Figure 12: The Cognitive Process representing the minimum components required for a system to deliver the suggested

conversational explanation and to facilitate Broad-XAI. It consists of four components: Perception, Merkwelt Model,

Actor’s Model and a Behaviour Model. Arrows between components represent the direction of inﬂuence that is required

by the level of explanation, indicated by the alphanumeric character in the arrow, along with a matching colour to previous

ﬁgures.

5.2. An Underlying AI-based Cognitive Process

The cognitive process can be regarded as the thinking, problem-solving and doing portion of the brain. In this

paper, we have regarded this cognitive process to be the machine intelligence that governs an agent’s behaviour and

is to be explained by the social process 5.1. Throughout, this paper has assumed the existence of a cognitive process

based on an underlying model of AI. This is not necessarily a common model, although the components in various

forms are often used. This section will brieﬂy discuss this assumed AI model and suggest approaches towards its

development. Figure 12 provides a detailed representation of this paper’s assumed AI model for the cognitive process.

This model suggests four components that work together to decide on an appropriate course of action, referred to as:

Perception, Merkwelt Model, Actor’s Model and a Behaviour Model. While not all of these components are required

for many AI systems, they are the minimum components of a system using all levels of explanation. Included in this

ﬁgure are arrows indicating the direction of inﬂuence of these components. It is these component inﬂuences that the

Zeroth-to-Nth levels (as indicated) aim to explain in the last section’s conversational process. While the yellow circles

indicate that the components themselves, and the explanations resulting from those components, are explained by the

Meta level.

Similar to agent-based approaches such as BDI-agents and Reinforcement Learning, the model assumes an external

environment that the agent perceives and processes to determine the preferred behaviour. Frequently, these are regarded

as a single process — that is to say, a single algorithm may be used to perform this input to decision mapping. For

as the basis for the Behaviour Model’s processing. For instance, the autonomous car must identify (either explicitly

or implicitly) each object in its vicinity prior to any behaviour based decision is processed. By separating these com-

ponents we aim to illustrate that the explanations of each of these are, while still purely reactive, aiming at providing

diﬀerent information. For instance, one common IML approach to explaining how a convolutional neural network

identiﬁed an object, such as a road sign, is through the use of saliency maps [20] — eﬀectively pointing a ﬁnger at

where it saw the object. Such an explanation is focusing on how the feature extraction layers of the network identiﬁed

the object. Often this feature extraction layer is a plugable component that can be used with additional models such

as decision trees [232] and random forests [233] for the subsequent mapping of behaviours. This allows additional

Reactive explanations detailing how the features aﬀected the choice of ﬁnal behaviour [108]. This process is shown

through the grey arrows through the centre of the model indicating these are all Reactive explanations.

All AI-systems have an implicit built-in set of motivations often deﬁned by the engineer deﬁning its success pa-

rameters or targets. This can be as simple as providing: a set of training examples; deciding on the number of clusters

to be found; or, deﬁning a reward structure. More recently many systems are incorporating an explicit attempt to model

their own internal motivations. This is particularly prevalent in systems where those internal motivations have the abil-

ity to be learnt or to switch between motivations during their operation, such as emotion inﬂuenced, multi-objective,

and sub-goal based approaches. Additionally, systems may explicitly provide a memory of past events that help guide

future decisions, such as: utilising memories of high rewarding experiences to bias decisions [234]; gates between

recurrent network iterations [235]; or, recalling past cases to guide a deliberative agent [236].

In this paper, we have introduced the idea of a Merkwelt Model to collectively refer to methods that have such

built-in motivations (either implicit or explicit) inﬂuencing their behaviours. The purpose of introducing this concept

is to highlight that these motivations are an important higher-level of reasoning that requires explicit explanation. In

Figure 12, we have indicated with the red arrow that the Merkwelt Model inﬂuences the behaviour of an agent and that

a Disposition explanation is required to detail the agent’s motivations. For example, the autonomous car’s motivation

to be in the right lane based on a sub-goal to exit the road at the approaching exit. Additionally, the cognitive process

illustrates that the Merkwelt Model also inﬂuences the model of the agent’s future expectations of other actors in

the environment. This is shown as a green arrow indicating the Nth-order explanation, as discussed in section 4.4,

illustrating the agent’s motivation is actually to alter other actors’ future behaviours.

The fourth component at the bottom quarter of Figure 12 is the inclusion of an Actor’s model that is required to

provide details of higher-level reasoning about external actors. Simple autonomous systems build a reactive model to

the environment and this may implicitly model external actors and how they may respond to the agent’s behaviours.

For example, a chess playing system may assume the opposition will select its best choice. However, when operating

in complex mixed domains an agent needs to explicitly model the external actor’s expected behaviour based on a model

of its motivations. For example, using pedestrian prediction to determine their intentions and future behaviour [146–

148]. These predictions directly inﬂuence the agent’s choice of behaviour, and therefore, often may be required to

explain its decisions. This is represented by the purple arrow between the Actor’s model and the agent’s Behaviour

model describing the need for a Social explanation.

Similarly, the predicted behaviour of an actor may also inﬂuence the agent’s own internal motivations represented

by its Merkwelt Model. For instance, if the prediction is that the actor will move to be in their future path then this

may cause the agent to switch focus to a safety-based avoidance objective [101, 237]. Explaining changes to an agent’s

internal beliefs and motivations is a recognition of the cultural expectations on the agent and must be explained with

Cultural explanations as indicated by the green arrow from the Actor’s Model to the agent’s Merkwelt Model. In

reality, the changes to the Merkwelt Model from the Actor’s Model and vice versa do not occur directly as indicated

by Figure 12, but instead are changed by observing the environment and inferring. This is indicated by the inclusion

of green arrows between Perception and the Merkwelt and Actor’s Models.

Finally, each of the explanations generated to detail the inﬂuences of components, plus the components themselves,

also need to explain how they generated their results directly. These represent Meta-explanations, which are shown

as yellow circles on each component and implied within each transition arrow. As discussed in section 4.5, these are

Level/Type of

Generalisation Required

Address

Assumptions

Explainee’s

Merkwelt

Quiescence

Model

Interpret

Explanation

Question

Behaviour

Perception

Model

th

N .

order Exp.

Actor’s

nd

Order Exp.

2

Model

st

order Exp.

1

th

0

order Exp.

Generalised Causes and

counterfactuals

Cognitive Process

Social Process

Figure 13: Explainer’s Explanation Process, elaborating on Figure 2, shows how the cognitive process links to the social

process. This shows that the cognitive process determines the agent’s behaviour based on the environment the agent ﬁnds

itself. The social process uses the levels of explanation to determine which of the components’ inﬂuence should be used

when providing an explanation.

5.3. The Explainer’s Explanation Process and Current Broad-XAI Implementations

In the previous sub-sections, we have discussed how the agent determines the level of explanation required to

resolve a user query through an iterative conversational process. This conversational process was designed as a social

interface with a human’s social process. Secondly, we outlined the cognitive process required of a machine intelligence

to be able to support such a Broad-XAI conversational process. Figure 13 now brings these processes together to detail

the explainer’s ﬁnal explanation process. Here we illustrate that the agent’s cognitive process, governing its behaviour,

links directly to the levels of explainability introduced in this paper. When explaining the agent’s behaviour the social

process uses the iterative conversation process, described in sub-section 5.1, which in turn maps the required level to

the appropriate component and how it is inﬂuencing the agent’s behaviour and its internal motivations and predictions.

For example, in explaining the cause of the accident in our motivating example a user may ask ‘why didn’t the car

give way to the pedestrian pushing a bicycle’? In answering this question, as discussed in sub-section 5.1, the system

should explain at the highest relevant level, which in this case would be the agent’s expectation that the pedestrian

would give way. This explanation is drawing from the Actor Model’s inﬂuence over the agent’s behaviour, resulting

in the explanation that the car ‘believed the person would expect the car to continue’. In the event that this explanation

does not satisfy the explainee the conversational process may require an additional explanation of what the agent’s

While this is the ﬁrst philosophical drawing together of these ideas and development of levels of explainability there

have already been some attempts to build such Broad-XAI approaches — although they have not been referred to as

such. In simple forms, there have been attempts to unify Reactive explanations of perception with behaviour [108, 238]

and attempts suggested extensions that combine these Merkwelt models [109, 239, 240].

More extensive coverage of this approach is represented using generic explanation facilities. These approaches

operate entirely external to the cognitive process and build a model based on their observation of an agent’s behaviour.

It is based on this observed model that it provides an explanation — hence these system tend use grey box models

to predict the agent’s behaviour. Two popular approaches are the Local Interpretable Model-Agnostic Explanations

(LIME) [155] and Black Box Explanations through Transparent Approximations (BETA) [241]. These system do not

provide an implementation anywhere near the breadth suggested as required in this paper but do often provide details

of both Reactive and Disposition explanations [156, 157].

One particularly notable extension to LIME attempts to implement a model that carries many similarities to the

approach described in this paper. Neerincx et al. [242] describes three phases of an explanation: "-generation, "-

communication, and "-reception. The "-generation phase separates the explanation of perception, behaviour and

cognition creation (that we have called Merkwelt) which align with our lower two levels of explainability. The "-

communication component standardises the communication with adaptive and interactive ontologies to be presented

to the user, while the "-reception measures the response of the user to the explanation aligning to our utility based on

quiescence. These three components align to the three components identiﬁed in our model of a conversational process

in section 5.1. While this work provides little philosophical or theoretical foundation to their approach and only covers

the lowest two levels of explanation, it does present a clear indication that the approach suggested in this paper has

potential to be implemented in real-world applications.

6. Conclusion

As AI is increasingly being integrated into society, research ensuring people can trust these systems has seen the

emergence of eXplainable AI (XAI) as an important topic. This is particularly the case with the utilisation of black-

box machine learning methods for which people have no intuition or understanding. A signiﬁcant amount of XAI

research, however, has been closely aligned to individual learning algorithms and have paid little attention to the need

of regular people to be able to interact and understand those explanations. This paper has argued that, in order to provide

acceptable and trusted explanations, a system must continually determine an explainee’s contextual position through an

interactive process, typically referred to as a conversation. Through this conversation, the agent progressively moves

the explainee towards the point of quiescence — a state of being quiet, where they have understood and accepted the

decision.

Furthermore, this paper has argued that AI cognition is structurally diﬀerent from that of humans and that XAI

systems need to go beyond cognitive interpretation alone and instead integrate technologies to achieve human-aligned

conversational explanations. This paper’s thesis has been that through the integration of technologies a Broad-XAI

system can be designed to provide conversational explanations that better align the cognitive process of an AI system

to that of people. This was accomplished by deﬁning a set of levels for XAI explanation, Figure 4, that cognitively

aligns to the AI process. These levels were designed to speciﬁcally map to the psychological model of the human

social process to address the three key areas of human explanation: contrastive explanation, attribution theory and

explanation selection. There were ﬁve levels of explainability identiﬁed in this paper: Zero-order (Reactive); First-

th

order (Disposition); Second-order (Social); N -order (Cultural) explanations; and, Meta (Reﬂective) Explanation.

These levels were drawn from ethology’s work in explaining animal intentionality.

Zero-order (Reactive) explanations focus on interpreting an individual decision based on the direct information

provided to it for that decision. This level is interested in the automatic reaction made based on an input set of features

and does not provide an explanation for any other factors such as an agent’s motivations or memory of past events.

For instance, a Convolutional Neural Network (CNN) classifying types of leaves may produce a heat map or identify

with the levels of XAI identiﬁed in the paper along with a selection of references identiﬁed.

Levels of Explanation

Fields of XAI Research Key papers

Reactive Disposition Social Cultural Reﬂective Conv

†

Transparent AI/ML ✓ ✓ [9–11]

†

Interpretable ML (IML) ✓ ✓ [12, 13, 89–96]

Explainable Agency ✓ [98]

Goal-driven XAI ✓ [103, 243]

‡

Memory-aware XAI ✓ [104–107, 109, 111, 112]

Emotion-aware XAI ✓ ✓ [113–115]

‡

Socially-aware XAI ✓ [156–158, 168–170]

‡

Culturally-aware XAI ✓ [192–197]

Object-level Explanation ✓ ✓ ✓ ✓ [220, 221]

∗

Deceptive Explanations ✓ [210–218]

Meta-Explanation ✓ [219–221, 244]

‡ ∗ ∗

Utility-driven XAI ✓ ✓ [201–203]

∗

Personalised Explanations ✓ [231, 245–247]

‡

Interactive XAI ✓ [223–226, 248]

‡

Broad-XAI ✓ ✓ ✓ ✓ ✓

∗

Field of research does not directly address this level but does represent a frequently included component.

†

Not used directly in this ﬁeld but can be used to partially explain the perception of other agents.

‡

Suggested future ﬁelds of research. Work listed was published under more general ﬁelds such as XAI or IML.

environment and other actors, and how that disposition motivated it towards a particular decision. For instance, a robot

that navigates towards its primary goal location decides to take the long way around a room to avoid bumping a vase

that occupied a narrow part of the more direct route. For a person to understand this behaviour the robot must explain

that the decision was based on a secondary objective to avoid damaging anything while moving around. This internal

motivation may also be inﬂuenced by: a memory of past events, such as the location of something it can no longer

directly perceive; or, internal parameters that are set or learnable, such as emotions. Research into explanations at this

level has been carried out in explainable agency, Goal-driven XAI, Memory-aware XAI, and Emotion-aware XAI.

Second-order (Social) explanations focus on decisions based on an agent’s awareness or belief of its own or other

actors’ mental states. This level builds on Disposition explanations and is aimed at explaining the agent’s interpretation

of what other actors might be thinking. In other words, if an agent’s decision is aﬀected by what it predicts others’

intentions might be then it should be able to explain how its decision was changed. This level is also referred to as

Social explanation as the agent is explaining a decision based on its social awareness and aligns with human reasoning

in social interactions. For example, an autonomous vehicle must predict the behaviour of other road users, cars, trucks,

pedestrians and animals and modify its behaviour accordingly. A Social explanation should be able to identify how the

prediction of other road users’ behaviour aﬀected its decisions. The raw interpretation of predictions of other actors is

frequently researched in existing Interpretable and Transparent ML, however, this only explains the prediction itself.

A true Social explanation will explain how that interpretation altered its decision. Early research in this area has been

discussed in both Emotion-aware XAI and Socially-aware XAI.

th

N -order (Cultural) explanations focus on decisions made by an agent on what it has determined is expected of it

culturally by other actors. This level requires an agent to not only model a prediction of other actors’ behaviour but

also models the expectations those actors may have about how the agent will behave. This level is also referred to

as Cultural explanations as it is concerned with the cultural expectations placed on an agent’s behaviour, which may

diﬀer in diﬀerent locations. For example, Cultural expectations on an agent about which side of a path an agent should

travel will change depending on whether it is on the Chinese mainland (right side) or Hong Kong (left side). The

provision of an explanation on how such cultural factors aﬀect an agent’s decision-making has only recently started to

Likewise, agents may be developed that are intentionally deceptive, such as negotiation based systems. Finally, in an

eﬀort to simplify and generalise an explanation crucial details may be omitted that inadvertently mislead. To address

this issue this paper also included a ﬁfth tangential type of explanation referred to as a Meta (Reﬂective) explanation.

A Meta-explanation is focused on explaining the process and factors that were used to generate, infer or select an

explanation.

Finally, this paper presented a model for explaining a decision or outcome through a recursive process of iterative

conversation. This process starts with a generalist assumption and progressively provides greater speciﬁcity until the

explainee reaches a point of acceptance, referred to as quiescence. We suggest that a mapping such as this allows

researchers and developers to facilitate trust and social acceptance of AI systems as they are increasingly integrated

into society through the provision of a contextually responsive system. In discussing each level we surveyed a range

of research previously applied in these areas and suggested future research directions to enable progress for each area.

Table 1 presents each of the ﬁelds discussed and maps them to the levels to which they may contribute. In this way,

this paper has pulled together all the disparate areas of XAI research and identiﬁed how each of these ﬁelds ﬁt together

into a broader context. In so doing, this paper aimed to promote new avenues of research into ﬁnding approaches for

Broad-XAI.

References

[1] E. H. Shortliﬀe, B. G. Buchanan, A model of inexact reasoning in medicine, Mathematical biosciences 23 (1975) 351–379.

[2] R. Davis, B. Buchanan, E. Shortliﬀe, Production rules as a representation for a knowledge-based consultation program, Artiﬁcial intelligence

8 (1977) 15–45.

[3] W. R. Swartout, XPLAIN: A system for creating and explaining expert consulting programs, Artiﬁcial intelligence 21 (1983) 285–325.

[4] B. Chandrasekaran, M. C. Tanner, J. R. Josephson, Explanation: the role of control strategies and deep models, Expert Systems: The User

Interface (1988) 219–247.

[5] C. Lacave, F. J. Díez, A review of explanation methods for bayesian networks, The Knowledge Engineering Review 17 (2002) 107–127.

[6] R. Andrews, J. Diederich, A. B. Tickle, Survey and critique of techniques for extracting rules from trained artiﬁcial neural networks,

Knowledge-based systems 8 (1995) 373–389.

[7] H. Cramer, V. Evers, S. Ramlal, M. Van Someren, L. Rutledge, N. Stash, L. Aroyo, B. Wielinga, The eﬀects of transparency on trust in and

acceptance of a content-based art recommender, User Modeling and User-Adapted Interaction 18 (2008) 455.

[8] M. Assad, D. J. Carmichael, J. Kay, B. Kummerfeld, PersonisAD: Distributed, active, scrutable model framework for context-aware services,

in: International Conference on Pervasive Computing, Springer, 2007, pp. 55–72.

[9] Y. Goyal, A. Mohapatra, D. Parikh, D. Batra, Towards transparent AI systems: Interpreting visual question answering models, arXiv preprint

arXiv:1608.08974 (2016).

[10] S. Wachter, B. Mittelstadt, L. Floridi, Transparent, explainable, and accountable AI for robotics (2017).

[11] C. Chao, M. Cakmak, A. L. Thomaz, Transparent active learning for robots, in: 2010 5th ACM/IEEE International Conference on Human-

Robot Interaction (HRI), IEEE, 2010, pp. 317–324.

[12] A. Abdul, J. Vermeulen, D. Wang, B. Y. Lim, M. Kankanhalli, Trends and trajectories for explainable, accountable and intelligible systems:

An HCI research agenda, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI’18, 2018.

[13] A. Adadi, M. Berrada, Peeking inside the black-box: A survey on explainable artiﬁcial intelligence (XAI), IEEE Access 6 (2018) 52138–

52160.

[14] C. Kuang, Can A.I. Be taught to explain itself?, 2017. URL: https://www.nytimes.com/2017/11/21/magazine/

can-ai-be-taught-to-explain-itself.html, accessed: 2019-10-06.

[15] D. Gunning, Explainable artiﬁcial intelligence (XAI), 2019. URL: https://www.darpa.mil/program/

explainable-artificial-intelligence, accessed: 2019-10-06.

[16] B. Goodman, S. Flaxman, European union regulations on algorithmic decision-making and a "right to explanation", arXiv preprint

arXiv:1606.08813 (2016).

[17] P. Voss, AGI Innovations (AGi3), 2019. URL: https://agiinnovations.com/, accessed: 2019-10-06.

[18] M. Hammond, Bonsai (AGi3), 2019. URL: https://bons.ai/, accessed: 2019-10-06.

[19] T. Miller, P. Howe, L. Sonenberg, Explainable AI: Beware of inmates running the asylum, in: IJCAI-17 Workshop on Explainable AI (XAI),

2017, p. 36.

[20] K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside convolutional networks: Visualising image classiﬁcation models and saliency maps,

arXiv preprint arXiv:1312.6034 (2013).

[21] M. D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: European conference on computer vision, Springer,

[25] D. H. Park, L. A. Hendricks, Z. Akata, A. Rohrbach, B. Schiele, T. Darrell, M. Rohrbach, Multimodal explanations: Justifying decisions

and pointing to the evidence, in: 31st IEEE Conference on Computer Vision and Pattern Recognition, 2018.

[26] E. Awad, S. Levine, M. Kleiman-Weiner, S. Dsouza, J. Tenenbaum, A. Shariﬀ, J.-F. Bonnefon, I. Rahwan, Blaming humans in autonomous

vehicle accidents: Shared responsibility across levels of automation, arXiv preprint arXiv:1803.07170 (2018).

[27] K. Gray, L. Young, A. Waytz, Mind perception is the essence of morality, Psychological Inquiry 23 (2012) 101–124.

[28] T. Miller, Explanation in artiﬁcial intelligence: Insights from the social sciences, arXiv preprint arXiv:1706.07269 (2017).

[29] D. J. Hilton, Mental models and causal explanation: Judgements of probable cause and explanatory relevance, Thinking & Reasoning 2

(1996) 273–308.

[30] B. R. Slugoski, M. Lalljee, R. Lamb, G. P. Ginsburg, Attribution in conversational context: Eﬀect of mutual knowledge on explanation-giving,

European Journal of Social Psychology 23 (1993) 219–238.

[31] T. Lombrozo, Simplicity and probability in causal explanation, Cognitive psychology 55 (2007) 232–257.

[32] P. Thagard, Explanatory coherence, Behavioral and brain sciences 12 (1989) 435–467.

[33] R. Dazeley, B. H. Kang, Epistemological approach to the process of practice, Minds and Machines 18 (2008) 547–567.

[34] J. von Uexküll, Mondes animaux et monde humain suivi de La théorie de la signiﬁcation, 1934.

[35] J. von Uexküll, A foray into the worlds of animals and humans: With a theory of meaning, volume 12, U of Minnesota Press, 2013.

[36] R. A. Brooks, Achieving Artiﬁcial Intelligence through Building Robots., Technical Report, MASSACHUSETTS INST OF TECH CAM-

BRIDGE ARTIFICIAL INTELLIGENCE LAB, 1986.

[37] R. A. Brooks, Intelligence without representation, Artiﬁcial intelligence 47 (1991) 139–159.

[38] W. Knight, What uber’s fatal accident could mean for the autonomous-car industry, 2018. URL: https://www.technologyreview.com/

s/610574/what-ubers-fatal-accident-could-mean-for-the-autonomous-car-industry/, accessed: 2019-10-06.

[39] NTSB, Preliminary report highway HWY18MH010, 2018. URL: https://www.ntsb.gov/investigations/AccidentReports/

Reports/HWY18MH010-prelim.pdf, accessed: 2019-10-06.

[40] A. Winﬁeld, Ethical standards in robotics and AI, Nature Electronics 2 (2019) 46.

[41] J. Woodward, Scientiﬁc explanation, 2017. URL: https://plato.stanford.edu/archives/fall2017/entries/

scientific-explanation/, accessed: 2019-10-06.

[42] E. o. Encyclopedia Britannica, Explanation, 2017. URL: https://www.britannica.com/topic/explanation, accessed: 2019-10-06.

[43] G. R. Mayes, Theories of explanation, 2001. URL: http://www.iep.utm.edu/explanat/, accessed: 2019-10-06.

[44] T. Mischel, Psychology and explanations of human behavior, Philosophy and Phenomenological Research 23 (1963) 578–594.

[45] G. Brown, Explaining, in: The handbook of communication skills, Routledge, 2006, pp. 205–238.

[46] H. P. Grice, Logic and conversation, 1975 (1975) 41–58.

[47] T. Lombrozo, The structure and function of explanations, Trends in Cognitive Sciences 10 (2006) 464 – 470.

[48] D. A. Wilkenfeld, T. Lombrozo, Inference to the best explanation (IBE) versus explaining for the best inference (EBI), Science & Education

24 (2015) 1059–1077.

[49] M. Schechtman, The narrative self (2011).

[50] Y. N. Harari, Homo Deus: A Brief History of Tomorrow, Harvill Secker, London, 2016.

[51] P. Salovey, J. Singer, The remembered self: Emotion and memory in personality, 1993.

[52] D. Kahneman, P. Egan, Thinking, fast and slow, volume 1, Farrar, Straus and Giroux New York, 2011.

[53] S. Mascarenhas, N. Degens, A. Paiva, R. Prada, G. J. Hofstede, A. Beulens, R. Aylett, Modeling culture in intelligent virtual agents,

Autonomous Agents and Multi-Agent Systems 30 (2016) 931–962.

[54] G. J. Hofstede, GRASP agents: social ﬁrst, intelligent later, AI & SOCIETY (2017) 1–9.

[55] A. Cawsey, User modelling in interactive explanations, User Modeling and User-Adapted Interaction 3 (1993) 221–247.

[56] G. I. Webb, M. J. Pazzani, D. Billsus, Machine learning for user modeling, User Modeling and User-Adapted Interaction 11 (2001) 19–29.

[57] S. C. Bakkes, P. H. Spronck, G. van Lankveld, Player behavioural modelling for video games, Entertainment Computing 3 (2012) 71 – 79.

[58] T. Pal, B. Chakraborty, J. Chakraborty, A survey of emotion recognition from handwritten script, International Journal of Innovative

Knowledge Concepts 7 (2019) 26–38.

[59] Y. Mehta, N. Majumder, A. Gelbukh, E. Cambria, Recent trends in deep learning based personality detection, arXiv preprint

arXiv:1908.03628 (2019).

[60] S. Rajan, P. Chenniappan, S. Devaraj, N. Madian, Facial expression recognition techniques: a comprehensive survey, IET Image Processing

13 (2019) 1031–1040.

[61] A. Chatterjee, G. Yasmin, Human emotion recognition from speech in audio physical features, in: Applications of Computing, Automation

and Wireless Systems in Electrical Engineering, Springer, 2019, pp. 817–824.

[62] C. Marechal, D. Mikołajewski, K. Tyburek, P. Prokopowicz, L. Bougueroua, C. Ancourt, K. Węgrzyn-Wolska, Survey on AI-based mul-

timodal methods for emotion detection, in: High-Performance Modelling and Simulation for Big Data Applications, Springer, 2019, pp.

307–324.

[63] F. Noroozi, D. Kaminska, C. Corneanu, T. Sapinski, S. Escalera, G. Anbarjafari, Survey on emotional body gesture recognition, IEEE

transactions on aﬀective computing (2018).

[64] K. S. Supriya, R. R. Reddy, Y. R. Devi, A survey on emotion’s recognition using internet of things, in: First International Conference on

[68] P. Biswas, P. Robinson, A brief survey on user modelling in HCI, in: Proc. of the International Conference on Intelligent Human Computer

Interaction (IHCI), volume 2010, 2010.

[69] O. Nocentini, L. Fiorini, G. Acerbi, A. Sorrentino, G. Mancioppi, F. Cavallo, A survey of behavioral models for social robots, Robotics 8

(2019) 54.

[70] M. Dutta, S. Mondal, S. Chakraborty, A. Chakraborty, A human intention detector - An application of sentiment analysis, in: Emerging

Technology in Modelling and Graphics, Springer, 2020, pp. 659–666.

[71] X. T. Truong, T. D. Ngo, Social interactive intention prediction and categorization, in: ICRA 2019 Workshop on MoRobAE-Mobile Robot

Assistants for the Elderly, Montreal Canada, May 20-24, 2019.

[72] H. C. Ravichandar, A. P. Dani, Human intention inference using expectation-maximization algorithm with online model learning, IEEE

Transactions on Automation Science and Engineering 14 (2016) 855–868.

[73] E. Pronin, T. Gilovich, L. Ross, Objectivity in the eye of the beholder: divergent perceptions of bias in self versus others., Psychological

review 111 (2004) 781.

[74] S. E. Toulmin, The Uses of Argument, 2 ed., Cambridge University Press, 2003. doi:10.1017/CBO9780511840005.

[75] B. Dickson, There’s a huge diﬀerence between ai and human intelligence—so let’s stop comparing them, 2018. URL: https://

bdtechtalks.com/2018/08/21/artificial-intelligence-vs-human-mind-brain/, accessed: 2020-11-01.

[76] S. Rosenthal, Why did the robot do that?, 2016. URL: https://insights.sei.cmu.edu/sei\_blog/2016/12/

why-did-the-robot-do-that.html, accessed: 2019-10-06.

[77] Explaining Decisions Made With AI, Alan Turing Institute and Information Commissioner’s Oﬃce (2020).

[78] E. Kazim, A. Koshiyama, Explaining decisions made with ai: A review of the co-badged guidance by the ico and the turing institute,

Available at SSRN 3656269 (2020).

[79] D. C. Dennett, The Intentional Stance, MIT Press, 1987.

[80] C. Stangor, Introduction to Psychology, Open Textbook Library, Flat World Knowledge, L.L.C., 2010.

[81] R. J. Hankinson, Cause and explanation in ancient Greek thought, Clarendon Press, 2001.

[82] A. Kass, D. Leake, Types of Explanations, Tech. Rep. ADA183253, DTIC Document, 1987.

[83] D. Marr, Vision: A computational investigation into the human representation and processing of visual information, 1982.

[84] T. Poggio, The levels of understanding framework, revised, Perception 41 (2012) 1017–1023.

[85] D. R. Griﬃn, The Question Of Animal Awareness: Evolutionary Continuity Of Mental Experience, Rockefeller University Press, San Mateo,

CA, 1976.

[86] D. L. Cheney, R. M. Seyfarth, How Monkeys See The World: Inside the mind of another species, University of Chicago Press, Chicago and

London, 1990.

[87] S. T. Mueller, R. R. Hoﬀman, W. Clancey, A. Emrey, G. Klein, Explanation in human-ai systems: A literature meta-review, synopsis of key

ideas and publications, and bibliography for explainable ai, arXiv preprint arXiv:1902.01876 (2019).

[88] F. Doshi-Velez, M. Kortz, R. Budish, C. Bavitz, S. Gershman, D. O’Brien, S. Schieber, J. Waldo, D. Weinberger, A. Wood, Accountability

of AI under the law: The role of explanation, arXiv preprint arXiv:1711.01134 (2017).

[89] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM

Computing Surveys (CSUR) 51 (2018) 93.

[90] F. Doshi-Velez, B. Kim, Towards a rigorous science of interpretable machine learning, arXiv preprint arXiv:1702.08608 (2017).

[91] B. Whitby, Artiﬁcial Intelligence, The Rosen Publishing Group, 2009.

[92] J. Huysmans, K. Dejaeger, C. Mues, J. Vanthienen, B. Baesens, An empirical evaluation of the comprehensibility of decision table, tree and

rule based predictive models, Decision Support Systems 51 (2011) 141–154.

[93] R. C. Fong, A. Vedaldi, Interpretable explanations of black boxes by meaningful perturbation, in: Proceedings of the IEEE International

Conference on Computer Vision, 2017, pp. 3429–3437.

[94] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, Y. Bengio, Show, attend and tell: Neural image caption generation

with visual attention, in: International conference on machine learning, 2015, pp. 2048–2057.

[95] L. M. Zintgraf, T. S. Cohen, T. Adel, M. Welling, Visualizing deep neural network decisions: Prediction diﬀerence analysis, arXiv preprint

arXiv:1702.04595 (2017).

[96] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, CoRR abs/1703.01365 (2017).

[97] M. Tegmark, Life 3.0: Being human in the age of artiﬁcial intelligence (2017).

[98] P. Langley, B. Meadows, M. Sridharan, D. Choi, Explainable agency for intelligent autonomous systems, in: Twenty-Ninth IAAI Conference,

2017.

[99] R. S. Sutton, A. G. Barto, Reinforcement Learning: An Introduction (Second Edition), Adaptive Computation and Machine Learning series,

A Bradford Book, London, 2018.

[100] A. S. Rao, M. P. Georgeﬀ, et al., BDI agents: from theory to practice., in: ICMAS, volume 95, 1995, pp. 312–319.

[101] P. Vamplew, R. Dazeley, C. Foale, S. Firmin, J. Mummery, Human-aligned artiﬁcial intelligence is a multiobjective problem, Ethics and

Information Technology 20 (2018) 27–40.

[102] P. Vamplew, C. Foale, R. Dazeley, Potential-based multiobjective reinforcement learning approaches to low-impact agents for ai safety

(submitted), Journal of Machine Learning Research (2020).

Agent Technology-Volume 03, IEEE Computer Society, 2011, pp. 17–20.

[106] M. Harbers, Explaining agent behavior in virtual training, Utrecht University, 2011.

[107] M. Harbers, K. van den Bosch, J.-J. Meyer, Design and evaluation of explainable BDI agents, in: 2010 IEEE/WIC/ACM International

Conference on Web Intelligence and Intelligent Agent Technology, volume 2, IEEE, 2010, pp. 125–132.

[108] P. Madumal, T. Miller, L. Sonenberg, F. Vetere, Explainable reinforcement learning through a causal lens, arXiv preprint arXiv:1905.10958

(2019).

[109] F. Cruz, R. Dazeley, P. Vamplew, Memory-based explainable reinforcement learning, 2019.

[110] O. Z. Khan, P. Poupart, J. P. Black, Minimal suﬃcient explanations for factored markov decision processes, in: Nineteenth International

Conference on Automated Planning and Scheduling, 2009.

[111] L. Arras, G. Montavon, K.-R. Müller, W. Samek, Explaining recurrent neural network predictions in sentiment analysis, arXiv preprint

arXiv:1706.07206 (2017).

[112] H. Bharadhwaj, S. Joshi, Explanations for temporal recommendations, KI-Künstliche Intelligenz 32 (2018) 267–272.

[113] F. Kaptein, J. Broekens, K. Hindriks, M. Neerincx, The role of emotion in self-explanations by cognitive agents, in: 2017 Seventh Interna-

tional Conference on Aﬀective Computing and Intelligent Interaction Workshops and Demos (ACIIW), IEEE, 2017, pp. 88–93.

[114] A. O. Rorty, Explaining emotions, The journal of philosophy 75 (1978) 139–161.

[115] P. O’Rorke, A. Ortony, Explaining emotions, Cognitive Science 18 (1994) 283–323.

[116] Y. Li, Deep reinforcement learning: An overview, arXiv preprint arXiv:1701.07274 (2017).

[117] K. Arulkumaran, M. P. Deisenroth, M. Brundage, A. A. Bharath, A brief survey of deep reinforcement learning, arXiv preprint

arXiv:1708.05866 (2017).

[118] Y. Duan, X. Chen, R. Houthooft, J. Schulman, P. Abbeel, Benchmarking deep reinforcement learning for continuous control, in: International

Conference on Machine Learning, 2016, pp. 1329–1338.

[119] M. Hossain, F. Sohel, M. F. Shiratuddin, H. Laga, A comprehensive survey of deep learning for image captioning, ACM Computing Surveys

(CSUR) 51 (2019) 118.

[120] D. Lee, H. Tang, J. O. Zhang, H. Xu, T. Darrell, P. Abbeel, Modular architecture for StarCraft II with Deep Reinforcement Learning, in:

Fourteenth Artiﬁcial Intelligence and Interactive Digital Entertainment Conference, 2018.

[121] L. A. Hendricks, Z. Akata, M. Rohrbach, J. Donahue, B. Schiele, T. Darrell, Generating visual explanations, in: European Conference on

Computer Vision, Springer, 2016, pp. 3–19.

[122] T. T. Nguyen, A multi-objective deep reinforcement learning framework, arXiv preprint arXiv:1803.02965 (2018).

[123] A. Abels, D. M. Roijers, T. Lenaerts, A. Nowé, D. Steckelmacher, Dynamic weights in multi-objective deep reinforcement learning, 2018.

arXiv:1809.07803.

[124] H. Mossalam, Y. M. Assael, D. M. Roijers, S. Whiteson, Multi-objective deep reinforcement learning, arXiv preprint arXiv:1610.02707

(2016).

[125] P. V. R. Ferreira, R. Paﬀenroth, A. M. Wyglinski, T. M. Hackett, S. G. Bilén, R. C. Reinhart, D. J. Mortensen, Multiobjective reinforcement

learning for cognitive satellite communications using deep neural network ensembles, IEEE Journal on Selected Areas in Communications

36 (2018) 1030–1041.

[126] A. M. Leslie, Pretense and representation: The origins of "theory of mind.", Psychological review 94 (1987) 412.

[127] H. Wimmer, J. Perner, Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children’s understanding

of deception, Cognition 13 (1983) 103–128.

[128] J. Holmes, Mentalisation: a key skill for psychiatrists and their patients, The British Journal of Psychiatry 193 (2008) 125–125.

[129] P. A. Lewis, A. Birch, A. Hall, R. I. Dunbar, Higher order intentionality tasks are cognitively more demanding, Social cognitive and aﬀective

neuroscience 12 (2017) 1063–1071.

[130] R. Saxe, Uniquely human social cognition, Current opinion in neurobiology 16 (2006) 235–239.

[131] M. Tomasello, J. Call, Primate cognition, Oxford University Press, USA, 1997.

[132] S. S. Adams, I. Arel, J. Bach, R. Coop, R. Furlan, B. Goertzel, J. S. Hall, A. V. Samsonovich, M. J. Scheutz, M. Schlesinger, S. C. Shapiro,

J. F. Sowa, Mapping the landscape of human-level artiﬁcial general intelligence, AI Magazine 33 (2012).

[133] B. Goertzel, Artiﬁcial general intelligence: Concept, state of the art, and future prospects, Journal of Artiﬁcial General Intelligence 5 (2014)

1–48.

[134] V. C. Müller, N. Bostrom, Future progress in artiﬁcial intelligence: A survey of expert opinion, in: Fundamental issues of artiﬁcial

intelligence, Springer, 2016, pp. 555–572.

[135] N. C. Rabinowitz, F. Perbet, H. F. Song, C. Zhang, S. Eslami, M. Botvinick, Machine theory of mind, arXiv preprint arXiv:1802.07740

(2018).

[136] S. Herath, M. Harandi, F. Porikli, Going deeper into action recognition: A survey, Image and vision computing 60 (2017) 4–21.

[137] C. Chen, R. Jafari, N. Kehtarnavaz, A survey of depth and inertial sensor fusion for human action recognition, Multimedia Tools and

Applications 76 (2017) 4405–4425.

[138] G. Cheng, Y. Wan, A. N. Saudagar, K. Namuduri, B. P. Buckles, Advances in human action recognition: A survey, arXiv preprint

arXiv:1501.05964 (2015).

[139] D. D. Dawn, S. H. Shaikh, A comprehensive survey of human action recognition with spatio-temporal interest point (stip) detector, The

[143] M. Al-Azzawi, R. Raeside, Modeling pedestrian walking speeds on sidewalks, Journal of Urban Planning and Development 133 (2007)

211–219.

[144] T. Gandhi, M. M. Trivedi, Pedestrian protection systems: Issues, survey, and challenges, IEEE Transactions on intelligent Transportation

systems 8 (2007) 413–430.

[145] T. Gandhi, M. M. Trivedi, Pedestrian collision avoidance systems: A survey of computer vision based recent studies, in: 2006 IEEE

Intelligent Transportation Systems Conference, IEEE, 2006, pp. 976–981.

[146] T. Hirakawa, T. Yamashita, T. Tamaki, H. Fujiyoshi, Survey on vision-based path prediction, in: International Conference on Distributed,

Ambient, and Pervasive Interactions, Springer, 2018, pp. 48–64.

[147] A. Rudenko, L. Palmieri, M. Herman, K. M. Kitani, D. M. Gavrila, K. O. Arras, Human motion trajectory prediction: A survey, arXiv

preprint arXiv:1905.06113 (2019).

[148] X. Wang, S. Zheng, R. Yang, B. Luo, J. Tang, Pedestrian attribute recognition: A survey, arXiv preprint arXiv:1901.07474 (2019).

[149] A. Mogadala, M. Kalimuthu, D. Klakow, Trends in integration of vision and language research: A survey of tasks, datasets, and methods,

arXiv preprint arXiv:1907.09358 (2019).

[150] N. Aafaq, A. Mian, W. Liu, S. Z. Gilani, M. Shah, Video description: a survey of methods, datasets and evaluation metrics, arXiv preprint

arXiv:1806.00186 (2018).

[151] D. Aineto, S. Jiménez, E. Onaindia, M. Ramírez, Model recognition as planning, in: Proceedings of the International Conference on

Automated Planning and Scheduling, volume 29, 2019, pp. 13–21.

[152] H. Xu, B. Li, V. Ramanishka, L. Sigal, K. Saenko, Joint event detection and description in continuous video streams, in: 2019 IEEE Winter

Conference on Applications of Computer Vision (WACV), IEEE, 2019, pp. 396–405.

[153] C. Roy, M. Shanbhag, M. Nourani, T. Rahman, S. Kabir, V. Gogate, N. Ruozzi, E. D. Ragan, Explainable activity recognition in videos., in:

IUI Workshops, 2019.

[154] N. Li, B. Liu, Z. Han, Y.-S. Liu, J. Fu, Emotion reinforced visual storytelling, in: Proceedings of the 2019 on International Conference on

Multimedia Retrieval, ACM, 2019, pp. 297–305.

[155] M. T. Ribeiro, S. Singh, C. Guestrin, Model-agnostic interpretability of machine learning, arXiv preprint arXiv:1606.05386 (2016).

[156] S. M. Mathews, Explainable artiﬁcial intelligence applications in NLP, biomedical, and malware classiﬁcation: A literature review, in:

Intelligent Computing-Proceedings of the Computing Conference, Springer, 2019, pp. 1269–1292.

[157] K. Weitz, D. Schiller, R. Schlagowski, T. Huber, E. André, Do you trust me?: Increasing user-trust by integrating virtual agents in explainable

AI interaction design, in: Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents, ACM, 2019, pp. 7–9.

[158] M. Hao, W. Cao, Z. Liu, M. Wu, Y. Yuan, Emotion regulation based on multi-objective weighted reinforcement learning for human-robot

interaction, in: 2019 12th Asian Control Conference (ASCC), IEEE, 2019, pp. 1402–1406.

[159] A. Khashman, A modiﬁed backpropagation learning algorithm with added emotional coeﬃcients, IEEE transactions on neural networks 19

(2008) 1896–1909.

[160] Y. Yang, Y. Wang, X. Yuan, F. Yin, Hybrid chaos optimization algorithm with artiﬁcial emotion, Applied Mathematics and Computation

218 (2012) 6585–6611.

[161] R. Thenius, P. Zahadat, T. Schmickl, EMANN-a model of emotions in an artiﬁcial neural network, in: Artiﬁcial Life Conference Proceedings

13, MIT Press, 2013, pp. 830–837.

[162] H. Yu, P. Yang, An emotion-based approach to reinforcement learning reward design, in: 2019 IEEE 16th International Conference on

Networking, Sensing and Control (ICNSC), IEEE, 2019, pp. 346–351.

[163] C. Balkenius, J. Morén, A computational model of context processing, in: 6th international conference on the simulation of adaptive

behaviour, 2000.

[164] A. Pentland, Socially aware, computation and communication, Computer 38 (2005) 33–40.

[165] H. Strömfelt, Y. Zhang, B. W. Schuller, Emotion-augmented machine learning: Overview of an emerging domain, in: 2017 Seventh

International Conference on Aﬀective Computing and Intelligent Interaction (ACII), IEEE, 2017, pp. 305–312.

[166] T. M. Moerland, J. Broekens, C. M. Jonker, Emotion in reinforcement learning agents and robots: a survey, Machine Learning 107 (2018)

443–480.

[167] D. Schuller, B. W. Schuller, The age of artiﬁcial emotional intelligence, Computer 51 (2018) 38–46.

[168] M. P. Sindlar, M. M. Dastani, F. Dignum, J.-J. C. Meyer, Mental state abduction of BDI-based agents, in: International Workshop on

Declarative Agent Languages and Technologies, Springer, 2008, pp. 161–178.

[169] M. P. Sindlar, M. M. Dastani, F. Dignum, J.-J. C. Meyer, Explaining and predicting the behavior of BDI-based agents in role-playing games,

in: International Workshop on Declarative Agent Languages and Technologies, Springer, 2009, pp. 174–191.

[170] M. Sindlar, M. Dastani, J.-J. Meyer, Programming mental state abduction, in: The 10th International Conference on Autonomous Agents

and Multiagent Systems-Volume 1, International Foundation for Autonomous Agents and Multiagent Systems, 2011, pp. 301–308.

[171] D. C. Dennett, Intentional systems in cognitive ethology: The “panglossian paradigm” defended, Behavioral and Brain Sciences 6 (1983)

343–355.

[172] R. Tuomela, The philosophy of sociality: The shared point of view, Oxford University Press, 2007.

[173] C. Adam, B. Gaudou, BDI agents in social simulations: a survey, The Knowledge Engineering Review 31 (2016) 207–238.

[174] J. S. Santos, J. O. Zahn, E. A. Silvestre, V. T. Silva, W. W. Vasconcelos, Detection and resolution of normative conﬂicts in multi-agent

[178] C. F. Camerer, Behavioral game theory: Experiments in strategic interaction, Princeton University Press, 2011.

[179] R. Suleiman, K. G. Troitzsch, N. Gilbert, Tools and techniques for social science simulation, Springer Science & Business Media, 2012.

[180] D. Silver, D. Hassabis, Alphago: Mastering the ancient game of go with machine learning, Research Blog 9 (2016).

[181] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot,

et al., Mastering the game of Go with deep neural networks and tree search, Nature 529 (2016) 484–489.

[182] C. Castelfranchi, Modelling social action for AI agents, Artiﬁcial intelligence 103 (1998) 157–182.

[183] R. Conte, C. Castelfranchi, et al., Cognitive and social action, Garland Science, 2016.

[184] I. Poggi, F. D’Errico, Cognitive modelling of human social signals., in: SSPW@ MM, 2010, pp. 21–26.

[185] K. Charalampous, I. Kostavelis, A. Gasteratos, Recent trends in social aware robot navigation: A survey, Robotics and Autonomous Systems

93 (2017) 85–104.

[186] Y. F. Chen, M. Everett, M. Liu, J. P. How, Socially aware motion planning with deep reinforcement learning, in: 2017 IEEE/RSJ International

Conference on Intelligent Robots and Systems (IROS), IEEE, 2017, pp. 1343–1350.

[187] R. Triebel, K. Arras, R. Alami, L. Beyer, S. Breuers, R. Chatila, M. Chetouani, D. Cremers, V. Evers, M. Fiore, et al., Spencer: A socially

aware service robot for passenger guidance and help in busy airports, in: Field and service robotics, Springer, 2016, pp. 607–622.

[188] B. Kim, J. Pineau, Socially adaptive path planning in human environments using inverse reinforcement learning, International Journal of

Social Robotics 8 (2016) 51–66.

[189] D. Vasquez, B. Okal, K. O. Arras, Inverse reinforcement learning algorithms and features for robot navigation in crowds: an experimental

comparison, in: 2014 IEEE/RSJ International Conference on Intelligent Robots and Systems, IEEE, 2014, pp. 1341–1346.

[190] H. Ritschel, Socially-aware reinforcement learning for personalized human-robot interaction, in: Proceedings of the 17th International

Conference on Autonomous Agents and MultiAgent Systems, International Foundation for Autonomous Agents and Multiagent Systems,

2018, pp. 1775–1777.

[191] C. Metz, How google’s AI viewed the move no human could understand, 2017. URL: https://www.wired.com/2016/03/

googles-ai-viewed-move-no-human-understand/, 2019-10-06.

[192] T. Kampik, J. C. Nieves, H. Lindgren, Explaining sympathetic actions of rational agents, in: International Workshop on Explainable,

Transparent Autonomous Agents and Multi-Agent Systems, Springer, 2019, pp. 59–76.

[193] T. Hellström, S. Bensch, Understandable robots-what, why, and how, Paladyn, Journal of Behavioral Robotics 9 (2018) 110–123.

[194] R. H. Wortham, A. Theodorou, Robot transparency, trust and utility, Connection Science 29 (2017) 242–248.

[195] A. D. Dragan, K. C. Lee, S. S. Srinivasa, Legibility and predictability of robot motion, in: Proceedings of the 8th ACM/IEEE international

conference on Human-robot interaction, IEEE Press, 2013, pp. 301–308.

[196] B. Fogg, G. Cueller, D. Danielson, Motivating, inﬂuencing, and persuading users: An introduction to captology, in: The human-computer

interaction handbook, CRC Press, 2007, pp. 159–172.

[197] S. Albert, Health captology–application of persuasive technologies to health care, E-Health: Current Status and Future Trends 106 (2004)

83.

[198] A. Nemery, E. Brangier, S. Kopp, How the use of persuasive criteria can enhance the response rate of a business web survey: one empirical

validation of the eight persuasive interaction criteria, Proceedings of Business and Information (2011).

[199] J. L. Conway Jr, Improving the Use of Mobile Medical Alert Devices in the Elderly, Ph.D. thesis, ProQuest Dissertations Publishing, 2019.

[200] T. Rist, M. Masoodian, Promoting sustainable energy consumption behavior through interactive data visualizations, Multimodal Technologies

and Interaction 3 (2019) 56.

[201] U. Ehsan, On design and evaluation of human-centered explainable AI systems, 2019.

[202] U. Ehsan, P. Tambwekar, L. Chan, B. Harrison, M. O. Riedl, Automated rationale generation: a technique for explainable AI and its eﬀects

on human perceptions, in: Proceedings of the 24th International Conference on Intelligent User Interfaces, ACM, 2019, pp. 263–274.

[203] J. McLaughlin, Utility-directed presentation of simulation results, in: Proceedings of the Annual Symposium on Computer Application in

Medical Care, American Medical Informatics Association, 1988, p. 292.

[204] S. H. Kim, J. K. Kim, Explanation in a decision-theoretic consulting system: An axiomatic approach, Applied Artiﬁcial Intelligence an

International Journal 5 (1991) 393–409.

[205] R. P. Marinier, J. E. Laird, Emotion-driven reinforcement learning, in: Proceedings of the Annual Meeting of the Cognitive Science Society,

volume 30, 2008.

[206] R. Elliott, A model of emotion-driven choice, Journal of Marketing Management 14 (1998) 95–108.

[207] R. P. Marinier III, J. E. Laird, R. L. Lewis, A computational uniﬁcation of cognitive behavior and emotion, Cognitive Systems Research 10

(2009) 48–69.

[208] J. Hoey, T. Schröder, A. Alhothali, Aﬀect control processes: Intelligent aﬀective interaction using a partially observable markov decision

process, Artiﬁcial Intelligence 230 (2016) 134–172.

[209] S. C. Gadanho, J. Hallam, Robot learning driven by emotions, Adaptive Behavior 9 (2001) 42–64.

[210] B. Wright, M. Roberts, D. W. Aha, B. Brumback, When agents talk back: Rebellious explanations (2019).

[211] H. Van Ditmarsch, Dynamics of lying, Synthese 191 (2014) 745–777.

[212] C. Sakama, S. C. Tran, E. Pontelli, A logical formulation for negotiation among dishonest agents, in: Twenty-Second International Joint

Conference on Artiﬁcial Intelligence, 2011.

Programming and Nonmonotonic Reasoning, Springer, 2011, pp. 331–344.

[217] G. Zlotkin, J. S. Rosenschein, Incomplete information and deception in multi-agent negotiation., in: IJCAI, volume 91, 1991, pp. 225–231.

[218] C. Sakama, M. Caminada, A. Herzig, A formal account of dishonesty, Logic Journal of the IGPL 23 (2014) 259–294.

[219] J. Pitrat, et al., Meta-explanation in a constraint satisfaction solver, in: Information Processing and Management of Uncertainty in Knowledge-

based Systems IPMU, Citeseer, 2006, pp. 1118–1125.

[220] B. Galitsky, Formalizing theory of mind, in: Computational Autism, Springer, 2016, pp. 95–176.

[221] B. A. Galitsky, J. L. de la Rosa i Esteva, B. Kovalerchuk, Explanation versus meta-explanation: What makes a case more convincing, in:

FLAIRS Conference, 2010.

[222] C. Antaki, I. Leudar, Explaining in conversation: Towards an argument model, European Journal of Social Psychology 22 (1992) 181–194.

[223] K. A. Cerulo, Social interaction: Do non-humans count?, Sociology Compass 5 (2011) 775–791.

[224] W. S. R. Can, S. Do, J. Seibt, et al., Human-animal analogy in human-robot interaction, What Social Robots Can and Should Do: Proceedings

of Robophilosophy 2016/TRANSOR 2016 290 (2016) 360.

[225] A. Elder, Living with robots, The Philosophers’ Magazine (2019) 115–117.

[226] M. Coeckelbergh, Humans, animals, and robots: A phenomenological approach to human-robot relations, International Journal of Social

Robotics 3 (2011) 197–204.

[227] S. A. McLeod, Reductionism and holism, 2008. URL: https://www.simplypsychology.org/reductionism-holism.html, ac-

cessed: 2019-09-27.

[228] D. Walton, Examination dialogue: An argumentation framework for critically questioning an expert opinion, Journal of Pragmatics 38

(2006) 745–777.

[229] A. Arioua, M. Croitoru, Formalizing explanatory dialogues, in: International Conference on Scalable Uncertainty Management, Springer,

2015, pp. 282–297.

[230] D. Walton, A dialogue system speciﬁcation for explanation, Synthese 182 (2011) 349–374.

[231] J. Schneider, J. Handali, Personalized explanation for machine learning: A conceptualization, arXiv preprint arXiv:1901.00770 (2019).

[232] J. Li, Y. Wu, J. Zhao, L. Guan, C. Ye, T. Yang, Pedestrian detection with dilated convolution, region proposal network and boosted decision

trees, in: 2017 International Joint Conference on Neural Networks (IJCNN), IEEE, 2017, pp. 4052–4057.

[233] B. Yang, J. Yan, Z. Lei, S. Z. Li, Convolutional channel features, in: Proceedings of the IEEE international conference on computer vision,

2015, pp. 82–90.

[234] D. Ramani, A short survey on memory based reinforcement learning, arXiv preprint arXiv:1904.06736 (2019).

[235] S. Wang, J. Cao, P. Yu, Deep learning for spatio-temporal data mining: A survey, IEEE Transactions on Knowledge and Data Engineering

(2020).

[236] J. M. Corchado, R. Laza, Constructing deliberative agents with case-based reasoning technology, International Journal of Intelligent Systems

18 (2003) 1227–1241.

[237] P. Vamplew, R. Issabekov, R. Dazeley, C. Foale, A. Berry, T. Moore, D. Creighton, Steering approaches to pareto-optimal multiobjective

reinforcement learning, Neurocomputing 263 (2017) 26–38.

[238] A. Anderson, J. Dodge, A. Sadarangani, Z. Juozapaitis, E. Newman, J. Irvine, S. Chattopadhyay, M. Olson, A. Fern, M. Burnett, Mental

models of mere mortals with explanations of reinforcement learning, ACM Transactions on Interactive Intelligent Systems (TiiS) 10 (2020)

1–37.

[239] R. Sukkerd, R. Simmons, D. Garlan, Toward explainable multi-objective probabilistic planning, in: 2018 IEEE/ACM 4th International

Workshop on Software Engineering for Smart Cyber-Physical Systems (SEsCPS), IEEE, 2018, pp. 19–25.

[240] R. Sukkerd, R. Simmons, D. Garlan, Tradeoﬀ-focused contrastive explanation for mdp planning, arXiv preprint arXiv:2004.12960 (2020).

[241] H. Lakkaraju, E. Kamar, R. Caruana, J. Leskovec, Interpretable & explorable approximations of black box models, arXiv preprint

arXiv:1707.01154 (2017).

[242] M. A. Neerincx, J. van der Waa, F. Kaptein, J. van Diggelen, Using perceptual and cognitive explanations for enhanced human-agent team

performance, in: International Conference on Engineering Psychology and Cognitive Ergonomics, Springer, 2018, pp. 204–214.

[243] S. Anjomshoae, K. Främling, Intelligible explanations in intelligent systems (2019).

[244] W. Samek, K.-R. Müller, Towards explainable artiﬁcial intelligence, in: Explainable AI: Interpreting, Explaining and Visualizing Deep

Learning, Springer, 2019, pp. 5–22.

[245] C.-H. Tsai, P. Brusilovsky, Designing explanation interfaces for transparency and beyond., in: IUI Workshops, 2019.

[246] L. Quijano-Sanchez, C. Sauer, J. A. Recio-Garcia, B. Diaz-Agudo, Make it personal: a social explanation system applied to group recom-

mendations, Expert Systems with Applications 76 (2017) 36–48.

[247] A. Kirsch, Explain to whom? putting the user in the center of explainable AI, in: Proceedings of the First International Workshop on

Comprehensibility and Explanation in AI and ML 2017 co-located with 16th International Conference of the Italian Association for Artiﬁcial

Intelligence (AI\* IA 2017), 2017.

[248] K. Sokol, P. A. Flach, Glass-box: Explaining AI decisions with counterfactual statements through conversation with a voice-enabled virtual

assistant., 2018.