A Measure of Explanatory Effectiveness *

Towards a Formal Model of Explanation

Dylan R. Cope† and Peter McBurney

King's College London

Abstract In most conversations about explanation and AI, the recipient of the explanation (the *explainee*) is suspiciously absent, despite the problem being ultimately communicative in nature. We pose the problem 'explaining AI systems' in terms of a two-player cooperative game in which each agent seeks to maximise our proposed measure of *explanatory effectiveness*. This measure serves as a foundation for the automated assessment of explanations, in terms of the effects that any given action in the game has on the internal state of the explainee.

Keywords: Explanation · XAI · Explainee-centric · Artificial Intelligence · Algorithmic Information Theory · Dialogues

1 Introduction

The term *explanation* in artificial intelligence (AI) is often conflated with the concepts of *interpretability* and *explainable AI* (XAI), but there are important distinctions to be made. Miller (2019) defines interpretability and XAI as the process of building AI systems that humans can understand. In other words, by design, the AI's decision-making process is inherently transparent to a human. In contrast, explicitly explaining the decision-making to an arbitrary human is *explanation generation*. The latter is the subject of this paper. More specifically, we are working towards developing a formal framework for the automated generation and assessment of explanations.

Firstly, some key terminology: an *explanation* is generated through a dialectical interaction whereby one agent, the *explainer*, seeks to 'explain' some phenomenon, called the *explanandum*, to another agent, the *explainee*. In this article, we propose a novel measure of *explanatory effectiveness* that can be used to motivate artificial agents to generate good explanations (e.g. in the form of a reward signal), or to analyse the behaviours of existing communicating agents. We then define *explanation games* as cooperative games where two (or more) agents seek to maximise the effectiveness measure.

*Work done by DC is thanks to the UKRI Centre for Doctoral Training in Safe and Trusted AI (EPSRC Project EP/S023356/1). PM wishes to thank Simon Parsons and Elizabeth Sonenberg for discussions on these topics. DC would like to thank Alex Jackson and Nandi Schoots for helping to understand understanding.

†Correspondence: dylan.cope@kcl.ac.uk

2 Related Literature

Intepretability and XAI have received an abundance of recent attention (see Adadi & Berrada (2018) for a review). This is largely due to two factors; regulatory demands (UK Information Commissioner's Office 2019) and the emergence of highly-performant black-box models, such as deep neural networks, that are naturally inscrutable. However, the central crux of interpretability techniques is the need to define a fixed *interpretable domain* from which we can derive explanations. This presents two challenges: there are no formal procedures for determining if a given domain is interpretable; and a domain may be interpretable to some agents, but not others, or only within some contexts. Moving away from interpretability, the problem of explanation generation has a long history in AI (Mueller et al. 2019). To some, there is a sense in which generating explanations is the hallmark of intelligence itself (Schank 1984). To others, explanation is simply about building models – a process which is seen as merely instrumental to intelligent behaviour (Russell & Norvig 2010, Hutter 2005, Chaitin 2006).

In the philosophy of science the concept of explanation is posed in terms of generating descriptions of, or hypotheses regarding, latent phenomena. This has led to investigations of formal measures of *explanatory power*, with an early example being Popper's (1959) notion of the 'degree of corroboration'. This developed into a line of philosophers devising subjectivist definitions for capturing aspects of the 'goodness' of explanations or hypotheses (Lipton 2003, Glass 2002, Okasha 2000, Schupbach & Sprenger 2011). However, by the subjectivity of these measures they may only assess the degree to which one believes (or simply likes) an explanation, which is not necessarily correlated with the degree to which an explanation is actually true (or representative of the world).

Recently, calls have been made for the need for human-centred explanation (Kirsch 2017, Abdul et al. 2018). Yet, the framing of explanation generation as a cooperative problem between a human and machine dates back to the era of expert systems (Karsenty & Brezillon 1995, Johnson & Johnson 1993, Graesser et al. 1996). By articulating explanation as a formal dialogue, a related direction of investigation is *dialogue games* (McBurney & Parsons 2002). In particular, *information-seeking* (Walton & Krabbe 1995) and *education* (Sklar & Parsons 2004) dialogues are especially relevant. Sklar & Azhar (2018) conducted empirical research with a human-machine collaboration task where the agents participated in a dialogue and explanations were provided to a human based of an *argumentation framework* (Dung 1995).

3 What is Explanation?

In this work we treat explanatory processes as involving two agents — an explainer and an explainee — and the result is that the explainee *understands* the explanandum better by the end than they did at the start. We define 'an explanation' as any sequence of observations made by the explainee that leads to this result. Thus an explanation could be a piece of text or spoken language, but it could also be a diagram or a piece of interactive media.

With this we shift the problem onto formally defining a measure of an agent's 'understanding' of some arbitrary phenomenon. We approach the question in terms of four stances[1] towards comprehension, understanding as: (1) a *sensation* (Hume 1751); (2) *information compression* (Chaitin 2006, Zenil 2019, Maguire et al. 2016); (3) *performance capacity* (Turing 1950, Perkins 1993); or (4) *organised information* (Lakoff & Johnson 1980, Hofstadter & Sander 2012).

The sensation stance states that comprehension is a conscious experience — you understand something if you feel that you apprehend it. The compression stance says that understanding is the formulation of concise and accurate descriptions of phenomena. The performance stance argues that having information is not enough; you must also know how to use the information. The organised-information stance tells us that utilisation and compression are a byproduct of something more important; namely that the agent represents information in relation to their own conceptual framework. While each of the stances has issues of their own, combined they provide a persuasive account. In other words, if someone claims they understand something, they can use their information to do things, and their description of the phenomenon is concise, accurate, and grounded in other concepts that they understand, then it is hard to argue that they do not grasp the phenomenon.

4 Technical Background

4.1 Algorithmic Information Theory

Algorithmic Information Theory (AIT) is a view of information that takes a fundamentally computational approach (Solomonoff 1964, Kolmogorov 1968, Chaitin 1975). Formally, AIT is built on the notion of *Kolmogorov complexity*, denoted $K(x)$. $K(x)$ is defined as the length of the shortest program, p, on a Universal Turing Machine (UTM), U, that outputs x.

$$K(x) = \min_{p}\{|p| \ : \ U(p) = x\}, \text{ where } |p| \text{ measures the length of } p \tag{1}$$

The conditional Kolmogorov complexity, $K(x|y)$, is similarly defined by the length of the shortest program that produces x when given input y.

$$K(x|y) = \min_{p}\{|p| \ : \ U(yp) = x\} \tag{2}$$

Thus we can define a measure of mutual information:

$$I(x;y) = K(y) - K(y|x) \tag{3}$$

Unless otherwise specified, when we talk of the mutual information between two objects we will be referring to an application of Equation 3.

[1]These stances do not represent arguments defended by anyone in particular, but rather we are constructing them here as rhetorical tools to help decompose the problem.

4.2 Agents

In its most basic conception, 'an agent' is any system that makes observations and takes actions. For any agent $X^t \in \mathcal{X}$ at time t, we will denote that they make observations $o^t_X \in \mathcal{O}_X$ and take actions $a^t_X \in \mathcal{A}_X$. Another important factor in describing agents is their *internal state*. This phrase can refer to various aspects of an agent's cognition, but we are mostly interested in this object insofar as it stores information. Firstly, we will assume that an agent's internal state may fall into a variety of configurations, i.e. there exists a set of possible internal states for an agent, which we will denote $\mathcal{Z}_X$. Secondly, we will talk of information being 'encoded' in an agent's internal state. Given an object o, we will denote X's encoding of o as $\langle o \rangle_X$, where $\langle o \rangle_X \in \{p \ : \ U(p) = o\}$ for some UTM U. We will speak of the agent 'having' this encoding, or its internal state 'containing' this encoding. This is independent to how this is achieved, e.g. the agent's internal state may simply store a list of encodings, or multiple encodings may overlap in a distributed storage medium such as a neural network.

4.3 Universal Intelligence Theory

Universal Intelligence Theory (UIT), proposed by Legg & Hutter (2007), establishes a definition of machine intelligence based on algorithmic information theory and reinforcement learning. In order to meaningfully compare different performances over a potentially infinite number of time steps, the scope of possible environments is limited such that the sum of rewards (the return) is always less than one. We will refer this as the set of *bounded-test environments*. With this, the *universal intelligence* of an agent π is given by:

$$\Upsilon(\pi) = \sum_{\mu \in E} 2^{-K(\mu)} V^\pi_\mu \tag{4}$$

Where V^π_μ is the return that π achieves in environment μ.

4.4 Universal Artificial Intelligence

Consider a stochastic environment with dynamics described by a probability distribution $\mu(e_k|æ_{<k})$, where e_k is the percept (observation-reward tuple) given at time k, and $æ_{<k}$ is the action-percept history. In order to perform optimally, the agent in this environment must infer μ. This is known as the problem of induction. By combining Solomonoff induction (Solomonoff 1964) with von Neumann-Morgenstern rational decision-making (Morgenstern & von Neumann 1953), Hutter (2005) defines AIXI; an agent that chooses the best possible action at every time step given perfect inductive inference.

5 Formalising Understanding

5.1 Partitioning the Internal State

In order to devise a measure of understanding, we will start by defining partitions of the information in the internal state. These partitions are constructed with

respect to a given phenomenon $p \in \mathcal{P}$. There are four: the p-relevant information (all information related to p), the p-irrelevant (all information completely unrelated to p), the p-specific (the information that *only* relates to p), and the p-background information (all information that is not specifically related to p). In the following formal definitions we are using a particular notation that warrants explanation. As we have already established, z_X denotes the internal state of agent X. We denote p-relevant notation with a comma after the X followed by $*p$, $z_{X,*p}$. The star indicates that we are 'selecting' *all* of the information relevant to p, rather than only the information specific to p. When the star is omitted we are referring to to specific information regarding whatever follows the comma, e.g. $z_{X,p}$ is the p-specific information and $z_{X,\neg p}$ is the information specific to everything that is not p (the p-irrelevant information).

Definition 1 (p-Relevant Information). *Given a phenomenon $p \in \mathcal{P}$ and an agent X with internal state z_X, the p-relevant information $z_{X,*p} \in \mathcal{Z}_X$ is the object where $I(z_{X,*p}; p) = I(z_X; p)$ and $I(z_{X,*p}; z_X)$ is minimised, i.e. there exists no $z'_{X,*p}$ such that $I(z'_{X,*p}; p) = I(z_X; p)$ and $I(z'_{X,*p}; z_X) < I(z_{X,*p}; z_X)$.*

Definition 2 (p-Irrelevant Information). *Given a phenomenon $p \in \mathcal{P}$ and an agent X with internal state $z_X \in \mathcal{Z}_X$, the p-irrelevant information $z_{X,\neg p}$ is the object where $I(z_{X,\neg p}; p) = 0$ and $I(z_{X,\neg p}; z_X)$ is maximised, i.e. there exists no $z'_{X,\neg p}$ such that $I(z'_{X,\neg p}; p) = 0$ and $I(z'_{X,\neg p}; z_X) > I(z_{X,\neg p}; z_X)$.*

Definition 3 (p-Specific Information). *Given a phenomenon $p \in \mathcal{P}$ and an agent X with internal state $z_X \in \mathcal{Z}_X$, the p-specific information $z_{X,p}$ is the object where $I(z_{X,p}; p) > 0,\ I(z_{X,p}; p') = 0\ \forall p' \in \mathcal{P}, p' \neq p$ and the mutual information $I(z_{X,p}; z_X)$ is maximised, i.e. there exists no $z'_{X,p}$ such that $I(z'_{X,p}; p) > 0,\ I(z'_{X,p}, p') = 0\ \forall p' \in \mathcal{P}, p' \neq p$ and $I(z'_{X,p}; z_X) > I(z_{X,p}; z_X)$.*

Definition 4 (p-Background Information). *Given a phenomenon $p \in \mathcal{P}$ and an agent X with internal state $z_X \in \mathcal{Z}_X and p-specific information z_{X,p}$, the p-background information[2] $z_{X,*\neg p}$ is the object where $I(z_{X,p}; z_{X,*\neg p}) = 0$ and $I(z_{X,*\neg p}; z_X)$ is maximised, i.e. there exists no $z'_{X,*\neg p}$ such that $I(z_{X,p}; z'_{X,*\neg p}) = 0$ and $I(z'_{X,*\neg p}; z_X) > I(z_{X,*\neg p}; z_X)$.*

5.2 Information Compression

With these partitions we can define how compressed the p-relevant information is:

Definition 5 (p-Compression Factor). *Suppose a phenomenon $p \in \mathcal{P}$ and an agent X. The p-compression factor $c : \mathcal{X} \times \mathcal{P} \to (0, 1]$ is given as the ratio of the Kolmogorov complexity of the p-relevant information object to the size of the agent's encoding of that information:*

$$c(X, p) = \frac{K(z_{X,*p})}{|\langle z_{X,*p} \rangle_X|} \tag{5}$$

[2] By the notation $*\neg p$ we can see that this is 'everything relevant' to 'not p'.

5.3 Information Utilisation

Next, we will attempt to formalise the performance stance on understanding, i.e. we will try to define X's *information utilisation* of p. To do this, we will need to construct a set of 'fair tests of p' for X. We will start by noting: (1) A fair test for X should *require* X's background information; (2) a test of p should *require* information about p. We will use the formalisation of rational decision-making, AIXI, to 'benchmark' how information is utilised in a given environment. Unlike a typical test-taker, AIXI enters into an environment with no prior knowledge, and thus we must present any priors to AIXI as a part of its percept sequence. Therefore, to decide whether or not a given task meets the criteria outlined above we will construct a 'meta-task' for AIXI where relevant observations are prepended to the task.

Definition 6 ((X,p)-tests). *Given a phenomenon $p \in \mathcal{P}$ and an agent X with internal state $z_X \in \mathcal{Z}_X$, we start with the set of bounded-test environments E, we define the set of (X,p)-tests, $E_{X,p}$, as follows:*

$$E_{X,p} = \left\{ \mu \in E \ : \ V^{AIXI}_{(p,b)\circ\mu} = V^*_\mu > 0, \ V^{AIXI}_{(p)\circ\mu} = V^{AIXI}_{(b)\circ\mu} = V^{AIXI}_\mu = 0 \right\} \tag{6}$$

Where b is a shorthand for the p-background information $b = z_{X,*\neg p}$, and $\mathbf{x} \circ \mu$ denotes the construction of a new environment μ' such that:

$$\forall x_i \in \mathbf{x}, \ \forall a_{<i}, \ \mu'((x_i, 0) \mid a_{<i}) = 1 \tag{7}$$

$$\forall k > |\mathbf{x}|, \ \mu'(e_k \mid a_{<k}) = \mu(e_k \mid a_{j\ldots k}), \ \textit{where } j = |\mathbf{x}| \tag{8}$$

It is worth noting why we are using only the p-background information and not the agent's entire internal state as required prior knowledge. This is because if the agent knows anything about p then AIXI would be able to use the information encoded in the internal state to pass the test when only given b. We want AIXI to only get information about p from p itself so that we can strictly outline the criteria above.

Using the set of fair tests for X, we can define a measure of *information utilisation* by measuring the agent's intelligence across these environments. This is an adaptation of Hutter's (2005) measure of intelligence (Equation 4).

Definition 7 (p-Utilisation). *Given an agent X and phenomenon p, the p-utilisation $\Upsilon_p : \mathcal{X} \to [0,1]$ is defined:*

$$\Upsilon_p(X) = \sum_{\mu \in E_{X,p}} 2^{-K(\mu)} V^X_\mu \tag{9}$$

5.4 Information Integration

With the definitions we have constructed here, we can also introduce a measure of how 'integrated' the p-relevant information is.

Definition 8 (p-Integration). *Suppose we have a phenomenon $p \in \mathcal{P}$, and an agent $X \in \mathcal{X}$ with p-relevant information $z_{X,*p}$ and p-specific information $z_{X,p}$. The p-integration, $\phi : \mathcal{X} \times \mathcal{P} \to [0, 1)$, is defined,*

$$\phi(X, p) = \tanh\left(\frac{|\langle z_{X,*p}\rangle_X|}{|\langle z_{X,p}\rangle_X|} - 1\right) \tag{10}$$

As the p-relevant information will always be larger-than or equal to the p-specific information ($|\langle z_{X,*p}\rangle_X| \geq |\langle z_{X,p}\rangle_X|$), the ratio in this measure will equal 1 when all relevant information is specific. In this case, there is no relevant information that is used for anything else, i.e. the p-relevant information is not at all integrated into the rest of the internal state (or nothing else exists to integrate with). Conversely, the smaller the specific information gets, the more the relevant information must be sharing with encodings for other phenomena.

5.5 The Measure of Understanding

Finally we bring these ideas together to define our measure of understanding. The resulting measure is bounded by 0 and 1.

Definition 9 (Understanding). *Given an agent $X \in \mathcal{X}$ with internal state z_X and phenomenon $p \in \mathcal{P}$, the measure of X's **understanding** of phenomenon p, $\kappa : \mathcal{X} \times \mathcal{P} \to [0, 1)$, is defined as:*

$$\kappa(X, p) = \frac{\hat{\kappa}(X, p) \cdot c(X, p) \cdot \phi(X, p) \cdot \Upsilon_p(X) \cdot I(z_X; p)}{K(p)} \tag{11}$$

Where $\hat{\kappa}(X, p) \in \{0, 1\}$ is X's self-reported understanding of p.

6 Explanation Games

With our measure of understanding, we are ready to define explanatory effectiveness:

Definition 10 (Explanatory Effectiveness). *The **effectiveness** of an explanation is the change in explainee's understanding of the explanandum $p \in \mathcal{P}$ over the course of the explanatory process. Formally, given an explainer agent A and an explainee agent B that interact over τ time steps, the explanatory effectiveness is a function $\xi : \mathcal{O}_B^* \times \mathcal{P} \to (-1, 1)$ defined as:*

$$\xi(\mathbf{o}_B,\ p) = \kappa(B^\tau,\ p) - \kappa(B^1,\ p) \tag{12}$$

Where B^t denotes B at time t and $\mathbf{o}_B$ is the sequence of observations that B made during the interaction.

Definition 11 (Explanation Game). *Suppose an explainer agent A, explainee agent B, and explanandum $p \in \mathcal{P}$. An **explanation game** $G = (A, B, p, \tau)$ is a cooperative finite sequential game with asymmetric information in which the participants seek to maximise $\xi(\mathbf{o}_B, p)$ over the course of τ time steps, where $\mathbf{o}_B$ is the sequence of all observations made by B.*

From these definitions, there are a few observations that we can make. Firstly, there is nothing to stop a game from having negative effectiveness, i.e. the explainee understands the phenomenon less after the 'explanation'. As κ is bounded by 0 and 1, ξ is bounded by -1 and 1. Secondly, there is no necessary link between effectiveness and the explainee's beliefs regarding their own understanding. It is possible for the explainee to believe that the explanation was more effective than it was (e.g. $\hat{\kappa}(X, p) = 1$, but $I(z_X; p) = 0$). Thirdly, we can use this notion to discuss the motivation of the explainer. It makes sense to consider an agent as an explainer, rather than a deceiver, only if they expect the sign of the ξ to be positive. Finally, it is worth noting that this measure changes according to time in which we choose to record it. The explainer may start out strong and increase the explainee's understanding of the explanandum, but then say something that leads to confusion.

7 Discussion

In this paper we have presented a formal model for assessing the the 'explanatory effectiveness' ξ of a dialectical process between two agents. We used this to define *explanation games* in which participants seek to maximise ξ. Along the way we used AIT and UIT to develop a measure of an agent's 'understanding' of a given phenomenon p. This involved partitioning the information in the agent's mental state into four objects relative to p; the p-relevant, p-irrelevant, p-specific, and p-background information. We used these to define the p-compression factor (how compressed the agent's representation of p is), p-integration (what proportion of the representation is only encoding for p), and the p-utilisation. For the last of these we needed to construct a set of 'fair tests', i.e. a set of environments that would rely on both knowledge of p and the agent's background knowledge to solve. We find these environments by asking: "Could AIXI solve this environment when given this information?". However, it should not be taken for granted that this is the right question to ask, and thus we should study this space of environments more precisely to see if it includes unfair tests or leaves out potential fair tests.

Future work should investigate the trustworthiness of explanations generated in our framework, as we have made the implicit assumption that if an agent understands something they can assess whether or not they trust it. One direction to look in is the implications of explainees with limited capacities, i.e. either time/space complexity constraints, or explainees who are biased in particular ways. Additionally, the assumption that explanation games are always cooperative should be challenged, as in many real situations participants may have conflicting or ulterior agendas. For both the cooperative and non-cooperative case a useful research project will be to articulate rules for the dialogue game between explainer and explainee (McBurney & Parsons 2002) and to develop strategies for each player, given their goals. Finally, as K and AIXI are not computable, alternatives for these components (for the purposes of this framework) should be devised and studied.

Bibliography

Abdul, A., Vermeulen, J., Wang, D., Lim, B. Y. & Kanhanhalli, M. (2018), Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda, *in* 'Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems', pp. 1–18.

Adadi, A. & Berrada, M. (2018), 'Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)', *IEEE Access* **6**, 52138–52160.

Chaitin, G. (2006), 'The Limits of Reason', *Scientific American* .

Chaitin, G. J. (1975), 'A Theory of Program Size Formally Identical to Information Theory', *Journal of the Association for Computing Machinery* **22**, 329–340.

Dung, P. M. (1995), 'On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games', *Artificial Intelligence* **77**(2), 321–357.

Glass, D. H. (2002), Coherence, Explanation, and Bayesian networks, *in* 'Lecture Notes in Artificial Intelligence', Vol. 2464, Springer Verlag, pp. 177–182.

Graesser, A. C., Baggett, W. & Williams, K. (1996), 'Question-driven Explanatory Reasoning', *Applied Cognitive Psychology* **10**(7), 17–31.

Hofstadter, D. R. & Sander, E. (2012), *Surfaces and essences : Analogy as the fuel and fire of thinking*, Basic Books.

Hume, D. (1751), *An Enquiry Concerning Human Understanding*.

Hutter, M. (2005), *Universal Artificial Intelligence*, Texts in Theoretical Computer Science, Springer Berlin Heidelberg, Berlin, Heidelberg.

Johnson, H. & Johnson, P. (1993), 'Explanation facilities and interactive systems', *International Conference on Intelligent User Interfaces* pp. 159–166.

Karsenty, L. & Brezillon, P. J. (1995), 'Cooperative problem solving and explanation', *Int. J. Expert Systems with Applications* **8**(4), 445–462.

Kirsch, A. (2017), 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)'.

Kolmogorov, A. N. (1968), 'Three approaches to the quantitative definition of information', *International Journal of Computer Mathematics* **2**(1-4), 157–168.

Lakoff, G. & Johnson, M. (1980), *Metaphors We Live By*, University of Chicago Press.

Legg, S. & Hutter, M. (2007), 'Universal intelligence: A definition of machine intelligence', *Minds and Machines* **17**(4), 391–444.

Lipton, P. (2003), *Inference to the Best Explanation*, 2 edn, Routledge (2003).

Maguire, P., Moser, P. & Maguire, R. (2016), 'Understanding Consciousness as Data Compression', *Journal of Cognitive Science* **17**, 63–94.

McBurney, P. & Parsons, S. (2002), 'Games That Agents Play: A Formal Framework for Dialogues between Autonomous Agents', *Journal of Logic, Language, and Information* **11**, 315–334.

Miller, T. (2019), 'Explanation in Artificial Intelligence: Insights from the Social Sciences', *Artificial Intelligence* **267**, 1–38.

Morgenstern, O. & von Neumann, J. (1953), *Theory of games and economic behavior*, Princeton University Press.

Mueller, S. T., Hoffman, R. R., Clancey, W., Emrey, A. & Klein, G. (2019), Explanation in Human-AI Systems: A Literature Meta-Review, Synopsis of Key Ideas and Publications, and Bibliography for Explainable AI, Technical report, DARPA.

Okasha, S. (2000), 'Van Fraassen's critique of inference to the best explanation', *Studies in History and Philosophy of Science Part A* **31**(4), 691–710.

Perkins, D. (1993), What is Understanding?, *in* 'Teaching for Understanding'.

Popper, K. (1959), *The Logic of Scientific Discovery*, Routledge Classics (2005), London and New York.

Russell, S. & Norvig, P. (2010), *Artificial Intelligence: A Modern Approach*, 3 edn, Pearson.

Schank, R. C. (1984), The Explanation Game, Technical report, Yale University.

Schupbach, J. N. & Sprenger, J. (2011), 'The Logic of Explanatory Power', *Philosophy of Science* **78**(1), 105–127.

Sklar, E. I. & Azhar, M. Q. (2018), Explanation through Argumentation, *in* 'Proceedings of the 6th International Conference on Human-Agent Interaction', Association for Computing Machinery (ACM), New York, NY, USA, pp. 277–285.

Sklar, E. & Parsons, S. (2004), Towards the application of argumentation-based dialogues for education, *in* 'Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems, 2004. AAMAS 2004.', pp. 1420–1421.

Solomonoff, R. J. (1964), 'A Formal Theory of Inductive Inference, Part 1', *Information and Control* **7**, 1–22.

Turing, A. M. (1950), 'Computing Machinery and Intelligence', *Mind* **LIX**(236), 433–460.

UK Information Commissioner's Office (2019), Guide to the GDPR, Technical report.

Walton, D. & Krabbe, E. C. W. (1995), Commitment in Dialogue: Basic Concepts of Interpersonal Reasoning, *in* 'SUNY series in Logic and Language', State University of New York Press.

Zenil, H. (2019), Compression is Comprehension, and the Unreasonable Effectiveness of Digital Computation in the Natural World, *in* S. Wuppuluri & F. A. Doria, eds, 'Unravelling Complexity', World Scientific, pp. 201–238.