Working Paper

2021/45/STR/DSC

Explainability as an Optimal Stopping Problem: Implications for Human-AI Interaction

Phanish Puranam
INSEAD, Phanish.puranam@insead.edu

Ilia Tsetlin
INSEAD, ilia.tsetlin@insead.edu

Version 1.0

Explainability is at the core of several phenomena related to the integration of information held by different actors (including the case involving machine intelligence and humans). We take as a premise that to explain something to someone requires finding a link between something they already know and the thing that is yet to be explained. We propose a formalization of this process as traversal across overlapping knowledge graphs, with the explainer facing an optimal stopping problem. This conceptualization helps to understand the barriers to explainability in terms of the relationships between the two graphs, and the costs and benefits of continuing search for a path that links what is to be explained to what is already in both graphs.

Electronic copy available at: http://ssrn.com/abstract=3914597

Acknowledgements: Puranam benefited from the Desmarais Fund to support the Organizations and Algorithms Project.

1. Introduction

The integration of specialized knowledge held by different actors requires a process that allows one actor to update their beliefs based on the other's. Such a process is core to diverse phenomena such as knowledge transfer, teaching, advice seeking and collaborative research; our "epistemic dependence" on each other is pervasive (Hardwig, 1985). Interest in explainability has been rekindled by recent developments in machine intelligence that can in principle improve human decision making if the human incorporates the analysis and recommendations of the machine. A major obstacle to the utilization of such recommendations has been the challenge of "explainability" that arises because of the complexity of the underlying models that machine intelligences use in order to make their predictions, which defy intuitive explanations in human-comprehensible terms.

We propose an approach to conceptualizing explanation as a graph traversal process. Our approach takes as a premise that to explain anything requires finding a link between something the "explainee" (to coin a term) already knows and the thing that is yet to be explained. This distinguishes explainability from acceptance based on trust alone (Hardin, 2002), or driven by conformity pressures (Asch, 1956) reinforced by cognitive dissonance reduction (Festinger, 1957). Rather, explainability requires furnishing ***reasons*** for accepting what the explainer proposes, and we assume those reasons take the form of linkages between what is already known and what is newly being offered.

Representing what is known by the explainer and explainee as graphs, we propose that the process of explanation can be understood as a traversal across overlapping knowledge graphs held respectively by the explainer and explainee, with the explainer facing an optimal stopping problem. This conceptualization helps to understand the barriers to explainability in terms of the relationships between the two graphs, and the costs and benefits of continuing search for a path that links what is to be explained to what is already in both graphs.

2. Related literature

Interest in explainability is currently most active in the field of applied computer science. The most sophisticated machine learning algorithms today rely extensively on non-linear function approximation. Explaining their working in an intuitive manner to humans is difficult (Lipton, 2018). This may create hesitations to adopt the algorithm's recommendations, as well as in some cases an overreliance on them (Bussone et al, 2015). In particular, machine learning applications in medicine confront the problem of explainability in a most visible manner, since the outcomes that depend on successful explanation involve human lives (see Tjoa and Guan, 2015 for a recent review of explainable AI applications in medicine).

In general, social influence as a basis for belief change requires that the recipient accept the information being offered. Incentive misalignments and communication failures (arising from the lack of a common language or channel noise) are two obvious impediments to such an acceptance. However, they do not exhaust the possibilities. Honest and motivated agents who can communicate using a common language and noise-free channels may nonetheless fail to explain themselves to each other because of "irreducible differences in perspectives" (Conner & Prahalad, 1996; Hong & Page, 2001).[1]

An important reason for this is that the absorption and assimilation of new information typically requires the ability to form connections with what is already known (Thagard, 2005). In an influential paper, Cohen and Levinthal (1990) argued that the dependence on prior knowledge for absorbing new information is the basis for understanding diverse phenomena such as the success (or failure) of knowledge transfer, why firms may invest in R&D even when the regime of appropriation is weak, and the extent of specialization and redundancy in

[1] This also relates to Thomas Kuhn's "incommensurability" hypothesis (though also see scepticism regarding this position from Anand, Larson and Mahoney (2020) who note that it might often be an excuse for intellectual turf protection).

expertise among the members of an organization. We draw on this insight to conceive of explanation as a search for knowledge elements that are known in common, in order to establish a link between what is to be explained and what the explainee already knows. Discovery of such a common element then enables the explainee to accept and absorb the novel element being offered by the explainer.

This acceptance has primarily a cognitive basis (i.e. to assimilate new information requires the ability to connect it to what is known), which is why we have described it as a “reason” to accept the explainer’s recommendation. However, once such a reason is found, it can also serve additionally as a credibility signal that operates through motivation: an explainee is more likely to accept a novel claim from an explainer when the explainer can demonstrate the claim’s connection to something that the explainee already accepts. The two channels may be intermingled as in the case of accepting a proof for a new mathematical claim based on commonly agreed axioms, and we do not distinguish further between “reasons” and “credibility signals”. Rather our focus is on the process of discovery of common knowledge elements and how that generates explainability.

3. An illustration

Before describing it in detail, to first illustrate the approach we take to conceptualize explainability and distinguish it from related concepts, consider a situation where a Department Chair is trying to decide whether to hire candidate A or candidate B for a faculty position. The Chair evaluates candidates on three dimensions – research, teaching and service. Based on their initial beliefs about the candidates along these dimensions, let us say the Chair prefers to hire A.

A colleague makes a recommendation (it could be to hire A or B) or offers additional information to the Chair. We define explainability – “eX” – of colleague’s recommendation as a property of the recommendation that causes the Chair to update their beliefs (including confidence in those beliefs) because the colleague has given sufficient reasons for it to be accepted. If the recommendation fails to cause the Chair to update their belief based on reasons, we say that the recommendation has not attained eX.

Note that we define eX as a property of the recommendation to rule out confounding with (relational) properties of colleague and Chair. For the same colleague-Chair combination, recommendation 1 may has more eX than recommendation 2 if the first is more likely to cause Chair to modify their beliefs based on reasons. This allows us to separate eX from things like power, trust, conformity or indoctrination – and can allow it to vary even if conditioned on the identity of explainer/explainee.

We propose that reasons are built on a listing of the steps – either logical or logistical, that the colleague undertook in order to reach the recommendation being made to the Chair. The key premise is that one or more of these steps must already be uncontroversial for the Chair, for the Chair to accept the recommendation. How many such steps need to be listed for the Chair to accept the information will depend on the specific colleague-Chair pairing, but for a given pairing, there will be a specific number of steps.

For instance, consider that the colleague offers a different set of scores on the same dimensions that the Chair considered and gives the steps that lead up to this conclusion (e.g. colleague has access to some private channels, those channels are trustworthy, the information could not be obtained except through private channels). Alternately, the colleague offers scores on a new set of dimensions that the Chair had not previously considered and gives arguments for why these dimensions matter (e.g. visibility in policy circles is important because the school is trying to build a reputation as a "force for good", therefore the Chair should weight this factor also, and here is the score for the candidates on these dimensions). In either case, our approach assumes that the Chair only accepts the colleague's recommendation or information (and eX has been achieved) if one or more of the steps leading up to the recommendation are already part of what the Chair believes.

4. A graph-theoretic formulation of eX

Consider a network with nodes that represent statements that are true (or alternately, not false), and links are relationships that imply mutual consistency. A link i←→j therefore reads as “the statement *i* is true is consistent with the statement *j* is true”. Consistency is symmetric and transitive. For instance, nodes can be “This candidate would do well in the INSEAD classroom”, “she did well when visiting LBS last year” and “LBS is very similar to INSEAD teaching environment”. These nodes can be connected with edges that represent mutual consistency.

This network is a part of objective reality (i.e. an immutable task environment). However, neither explainer nor explainee have this full network in mind. Rather they have **sub-graphs-** representations of portions of the network in their long-term memory (Thagard, 2005). These representations are incomplete – nodes that should exist do not. However, we assume that links (if present) are always accurate. Sub-graphs are incomplete because expanding their size (i.e. adding more nodes) is costly. Therefore, it must also be true that if eX arises, the benefit to explainee of adding a node must exceeds this cost. Finally, sub-graphs are private – they are mutually unknown, except through ***conversation***.

We assume a ***conversation*** proceeds as follows:

1. Explainer starts by listing a node in own network – node zero. This is what needs to be explained.
2. Explainee checks if the node is in their own network.
3. If

a. no, then explainer reveals another node that is consistent with previously revealed node in own sub-graph. Both new and previous node + their link is stored in *working memory* by explainee. Step 2 repeats. Conversation stops if all nodes are exhausted.

b. yes, then a ***bridge*** is created – a path between this node and all other nodes in working memory. The entire bridge is then added to the explainee's *long term memory* (i.e. sub-graph). Node zero has been "explained". This is an implication of transitivity - if i←→k ←→j, then i←→j.

Explainer and Explainee roles can also alternate within the conversation. In the context of interpretable AI, the explainer (an AI algorithm) will be rather passive, while the explainee (a human) will be the active questioner. Either a constraint on working memory for explainee (equivalent to constraint on how much explainee's sub-graph can potentially expand through this conversation) or a cost to explainer for revealing each node, or to receiver to check each offered node against own sub-graph will stop the conversation.

Note also that in this model, we are assuming away any challenge for explainee to accept a link – it is only nodes that need "explanation". One might need to go through many nodes but each link itself is unproblematic. An interesting implication is that while the repeated trials to discover common nodes are costly, it may be "better" to succeed on later rather than earlier trials. This is because nodes shown on all previous trials become part of the explainee's understanding. Frustrating conversations can be fruitful.

Within this framework, we can reconsider the illustration involving the recruitment problem discussed before. Figure 1 shows the states of the network before conversation between the Chair and colleague commences. Colours differentiate what is uniquely in each sub-graph, what is in common, and what is not in either. Node C is common to both Colleague and Chair sub-graphs. Figure 2 shows the state of network after conversation. The example discussed previously can be represented as a situation where a Colleague causes Chair to bring into existence a node “A” that was previously missing in Chair’s sub-graph. The Colleague does this by first revealing node A (which is not in Chair’s sub-graph), followed by C (which is), leading to the addition to Chair’s sub-graph the bridge A←→C. The “reason” is that colleague gets to a common node C, in 2 steps. If instead Colleague had gone for node B in step 2, explainability would not have been achieved. The reasons for a node inclusion (e.g., A went from being non-existent to existent in Chair’s sub-graph) ultimately comes from colleague being able to reach a common node C. Given the constraints on working memory/cost of node-link revelation, the relationship between the two sub-graphs ex ante (e.g. overlap, average cross-graph path lengths etc.) fully determines eX for this recommendation.

Alternatively, one can imagine that for every node offered by a colleague, Chair gets an unbiased estimate from the true underlying network of how many *more* steps will be required to hit a common node or validate an existing link. This seems a bit magical (this is perhaps the idea of expected “fruitfulness” of one approach vs. another to explain) but let’s assume for the moment it is possible. However, if such a signal exists, then we must fall back on an exogenous constraint on time/cost to decide when this signal looks unpromising. So again, the ex-ante relationship between the two sub- graphs will determine eX for a recommendation.

A formalization for the case of fully (internally) connected sub-graphs

While the problem, as formulated appears simple, there are substantial complexities involved in a general formalization of what a rational explainer might do when embarking on a conversation with the intention of attaining explainability.

To illustrate this, we consider a formulation of this traversal process for the special case where the sub-graphs are fully internally connected, using a simple urn-based model of sampling without replacement. Let the number of nodes in explainer’s network, that are not yet revealed to the explainee, be *n*. The explainer has a prior on the number of overlapping nodes *K*: from these *n* nodes, k_i nodes can be in the explainee's network with probability p_i', *i*=1,..,*m*. This formulation preserves the common intuition about optimal overlap between sub-graphs, which is a function of $\frac{K}{N}\left(1-\frac{K}{N}\right)$, where *N* is the total number of nodes in the explainer's network. Complete overlap (*K* = *N*) implies no value to the conversation, as there's no nodes in the explainer's network that are new to the explainee, and zero overlap (*K* = 0) implies no possibility of a fruitful conversation. Explainability eX is only worth pursuing *and* feasible when sub-graphs have partial overlap.

Denote $\mu_K' = \sum_{i=1}^{m} p_i' k_i$ – this is the expected number of overlapping nodes. Then the probability that eX will be achieved at the next step (i.e. after revealing one out of *n* nodes at random) is μ_K'/n. Also denote variance of the number of overlapping nodes by $V_K' = \sum_{i=1}^{m} p_i'(k_i - \mu_K')^2$.

If eX was not achieved after the first node has been shown, the explainer will update probabilities of k_i from p_i' to

$p_i'' = \frac{p_i'(n-k_i)/n}{1-\mu_K'/n} = p_i' \frac{n-k_i}{n-\mu_K'}$; The posterior mean for the number of overlapping nodes becomes

$$\mu_K'' = \sum_{i=1}^{m} p_i'' k_i = \sum_{i=1}^{m} p_i' \frac{n-k_i}{n-\mu_K'} k_i = \frac{n\mu_K' - (\mu_K')^2 - V_K'}{n-\mu_K'} = \mu_K' - \frac{V_K'}{n-\mu_K'}.$$

Now, the probability that eX will be achieved in the next step is $\mu_K''/(n-1)$.

This simple model formalizes the change in expectation of success after the first revealed node by the explainer is found not to be in the explainee's network. It delivers two insights.

1. The incentives to continue a conversation that aims to achieve explainability a) must decline with variability of the prior distribution on the extent of overlap of sub-graphs V_K' and b) d**e**creases with sub-graph size (n+1) for a given prior about expected overlap.

2. If the explainer faces a cost for revealing each node, the conversation can terminate despite $K > 0$ (and despite the net benefit to explainee of adding a node to long term memory always being positive). This illustrates why honest, motivated agents, with no communication difficulties (in the sense that communication may be costly but is error-free) may nonetheless be left with irreducible differences in perspectives- one may know something that it is simply not worth explaining to the other, *even if* in principle such explainability could be achieved (i.e. $K > 0$) (Conner and Prahalad, 1996).

The special case formalized above is quite limited. First, it does not incorporate incomplete connectivity within sub-graphs, which may lead to a path dependent process of node revelation by explainer, and which probably accounts for the failure of explainability in many if not most real-world conversations. A second limitation is that we have not yet considered "Gestalt" versions of eX, which might be illustrated as follows (Figure 3). In this version the conversation process outlined before continues with the following modification, beginning with the state of affairs as shown in Figure 1.

2': Receiver checks if the new node, *if* admitted into own network, validates any existing links through triadic closure.

The Chair asks (when shown A) – if I accept A, would it validate any of my existing links? In this case through triadic closure-the link CD could be validated (i.e. confidence in it could be enhanced). This is the canonical case for new theory in most sciences – a new fact A gains credibility because it is part of a "nomological network" (A, C, D) in which some ties were already accepted (in this case CD). For the colleague this might then in a next step lead to the addition of node D to their sub-graph. Both sides benefit from this conversation.

To accommodate such cases requires modelling priors over not just number of overlapping nodes but on overlapping paths in the sub-graphs, as well as the link strength. Formulating and solving the general case of optimal search on a graph can be very complicated even for modest sized graphs (Brown and Smith, 2013). Computational models of adaptively rational agents may offer a path to further useful insights (Puranam et al, 2015).

5. Implications

Some implications of this framework are sketched below to illustrate its potential fruitfulness.

i. ***Irreducible diversity in perspectives*** among honest, motivated agents who can communicate effectively may be quite common. As we have shown it arises even in a simple model with full intra-connectivity of sub-graphs. Without such connectivity, the possibility of path-dependence and cessation of conversation without attaining explainability may be greater.

ii. Explainability is an ***alternative to homophily*** as an explanation for similarity-based influence. Rather than invoke a motivation-based argument such as willingness or openness to influence based on similarity, explainability gives a cognitive/informational account (explanation!) for why social influence may be more effective among people who already hold similar beliefs (which in turn may lead to phenomena such as echo chambers or social fragmentation, e.g. Axelrod, 1995).

iii. Explainability offers a formal mechanism to approach knowledge integration processes in groups. For instance, it can justify the value of building ***transactive memory*** (Wegner, 1987) – knowledge of who knows what in a group – since it can generate sharper priors on the extent of overlap. It can also offer an alternative explanation for the phenomena of the ***curse of knowledge*** – for a given magnitude of absolute overlap, the larger the explainer's graph, the harder it is to attain eX, because of the greater possibility of taking long or dead-end paths. This is distinct from the usual mechanism that involves overestimation of overlap by the explainer.

iv. The graph-theoretic formulation, particularly with a gestalt like extension, offers a natural path to model ***encapsulation*/*compression.*** Through triadic closure, nodes can be encapsulated into a super-node. This may conserve capacity (and may amplify the willingness for the "reason" to be accepted).

v. Explainability-based acceptance of new facts is a force that increases internal connectivity among sub-graphs, whereas acceptance of other's opinions for reasons having to do with trust, conformity or cognitive dissonance will produce disjunct sub-graphs. This suggests a ***self-limiting*** nature to reason-based knowledge integration compared to a reason-free approach: sole reliance on reason-based integration produces internal coherence of what is known but may limit the possibilities of future expansion of what one can know.

vi. In the context of machine intelligence interacting with humans, more than a "friendly interface" is needed for successful knowledge transfer through explanation. The optimal way to connect new insights with the explainee's existing knowledge would be different for different categories of explainee's, and facilitating the exploration of different explanation paths, to reduce the ultimate cost of eX, could be a welcome addition to developments in the explainability literature in AI.

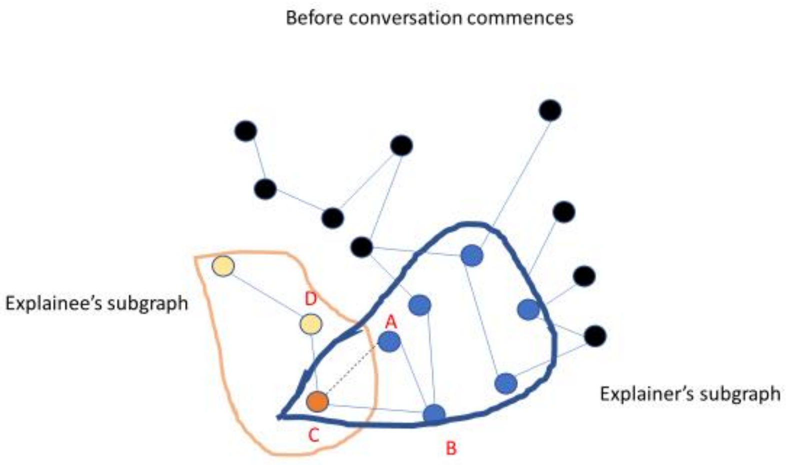

Figure 1: State of sub-graphs before conversation

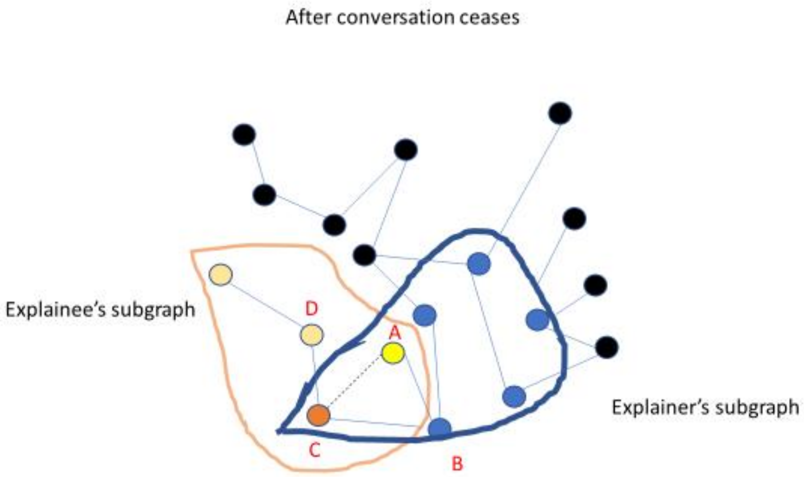

Figure 2: State of sub-graphs after conversation ceases

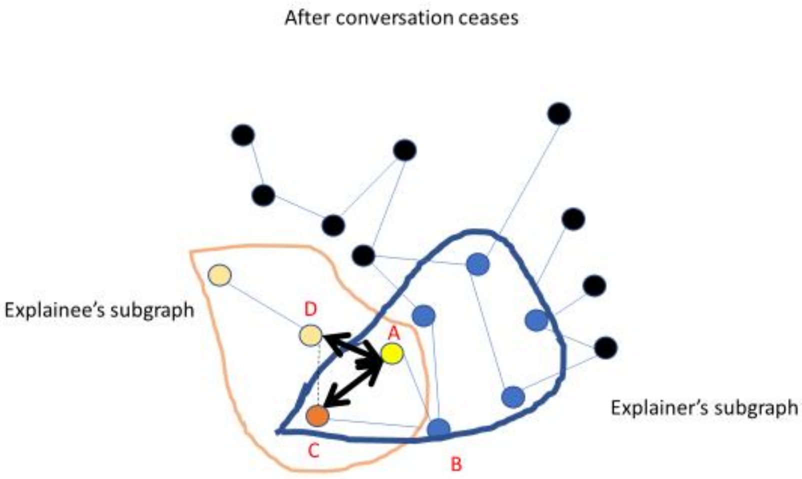

Figure 3: A “gestalt” version of explainability (after conversation ceases)

Bibliography

Anand, G., Larson, E.C. and Mahoney J.Y. (2020) Thomas Kuhn on paradigms, *Production and Operations Management* 29(7) : 650-1657

Asch, S. E. (1956) Studies of independence and conformity: I. A minority of one against a unanimous majority. *Psychological Monographs 70*, 1–70

Axelrod, R. (1995) The dissemination of culture: a model with local convergence and global polarization, *The Journal of Conflict Resolution* 41(2):203-226

Bussone, A., Simone, S. and O'Sullinav, D. (2015) The role of explanations on trust and reliance in clinical decision support systems, *International Conference on Healthcare Informatics*

Brown, D.B. and Smith, J.E. (2013) Optimal sequential exploration: Bandits, Clairvoyats and Wildcats *Operation Research* 61(3):644-665

Cohen, W. M., & Levinthal, D. A. (1990) Absorptive capacity: A new perspective on learning and innovation. *Administrative Science Quarterly 35*(1), 128-152.

Conner, K. R., & Prahalad, C. K. (1996) A Resource-based Theory of the Firm: Knowledge Versus Opportunism. *Organization Science 7*(5), 477-501. doi:10.1287/orsc.7.5.477

Festinger, L. (1957) *A Theory of Cognitive Dissonance*: Stanford University Press.

Hardin, R. (2002). *Trust and Trustworthiness.* In. New York, NY: Russell Sage Foundation.

Hardwig, J (1985) Epistemic dependence, *The journal of philosophy* 82(7):335-349

Hong, L. and Page, S.E. (2001) Problem Solving by Heterogenous Agents, *Journal of Economic Theory* 97, 123-163

Lipton, Z. C. (2018) The mythos of model interpretability. *Queue 16*(3), 31–57.

Puranam P, Stieglitz N, Osman M, Pillutla MM (2015) Modelling bounded rationality in organizations: Progress and prospects Acad. Management Ann. 9(1):337–392

Tjoa, E and Guan, C. (2015) A survey of Explainable Artifical Intellience (XAI): towards Medical XAI; Journalof Latex Class Files, Vol 14, No. 8

Thagard, P. (2005) *Mind*: MIT Press. Cambridge MA

Wegner, D. M. (1987) Transactive memory: A contemporary analysis of the group mind. *Theories of Group Behavior* 185-208.