Different "Intelligibility" for Different Folks

Yishan Zhou
Department of Cognitive Science
University of California, San Diego
San Diego, CA USA
yiz329@ucsd.edu

David Danks
Departments of Philosophy and Psychology
Carnegie Mellon University
Pittsburgh, PA USA
ddanks@cmu.edu

ABSTRACT

Many arguments have concluded that our autonomous technologies must be intelligible, interpretable, or explainable, even if that property comes at a performance cost. In this paper, we consider the reasons why some property like these might be valuable, we conclude that there is not simply one kind of 'intelligibility', but rather different types for different individuals and uses. In particular, different interests and goals require different types of intelligibility (or explanations, or other related notion). We thus provide a typography of 'intelligibility' that distinguishes various notions, and draw methodological conclusions about how autonomous technologies should be designed and deployed in different ways, depending on whose intelligibility is required.

KEYWORDS

Intelligibility; Explainability; Prediction algorithms

ACM Reference format:

Yishan Zhou and David Danks. 2020. Different "Intelligibility" for Different Folks. In *Proceedings of 2020 ACM Conference Artificial Intelligence, Ethics, and Society (AIES'20), Feb. 7-8, 2020, New York, NY.* ACM, NY, NY, USA, 6 pages. https://doi.org/10.1145/3375627.3375810.

1 Why Intelligibility?

Theoretical advances, accompanied by dramatic growth in computing power, have allowed the development of complex autonomous systems that operate on huge, high-dimensional datasets using sophisticated internal models and representations. The resulting systems exhibit diverse degrees not only of predictive power (given interventions or observations), but also of interpretability, intelligibility, and explainability. Unsurprisingly, there have been numerous recent papers that aim to understand the nature and value of these latter properties (e.g., [2, 3, 19, 20], and many more). For simplicity, we use 'intelligible' to refer generically to this collection of properties (including explainable, interpretable, and understandable) that all focus on what people can know or infer about a system. While there are many arguments about which of these properties is the "right" one, our focus here is on the general form that these properties should all have.

More intelligible AI certainly *seems* to be desirable, but one common concern is that we must tradeoff performance to achieve it (though see [15]). For example, decision trees are easily understood by people, but can have reduced predictive or classification power compared to more sophisticated methods based on deep neural networks, which can be relatively inscrutable to human understanding. In fact, there is a superficially plausible argument that we face a *necessary* tradeoff between intelligibility (and related notions) and performance. Suppose that we are trying to find the performance-optimal model within some model-class **M**. The model-class **I** of *intelligible* models is a (not necessarily proper) subset of **M**. Hence, the performance of the best model in **I** is upper-bounded by the best model in **M**. In practice, **I** is usually a *much* smaller set than **M**, and so (the simple argument continues) we almost always face a trade-off between intelligibility and model performance. Given the constant pressure to improve accuracy, the simple argument concludes that intelligibility is a disposable luxury.

We contend that intelligibility is a trickier matter than is presupposed by this argument (and many others). In particular, *intelligibility* is not a property of models (or model-classes) in isolation. In this paper, we first argue that there is no intrinsic value to intelligibility, or even accuracy. Rather, these properties or performance have value only to the extent that they help people to realize their goals and advance their interests. But that conclusion means that intelligibility is not a unitary notion: we need to consider different types of intelligibility based on their value to people in context with purposes or goals. Thankfully, there are not arbitrarily many such notions of intelligibility. Rather, we argue that there are only a few major types, and so we can have more focused criteria for intelligible autonomous systems, depending on their functions and contexts.

1.1 Different Needs for Different Folks

It is a banal observation that autonomous systems are designed for particular functions or uses. And the performance measures that we use for a system—accuracy, precision, recall, transfer, and so forth—are typically chosen precisely because they are informative

AIES '20, February 7–8, 2020, New York, NY, USA

ACM 978-1-4503-7110-0/20/02…$15.00
https://doi.org/10.1145/3375627.3375810

about the likely ability of the system to satisfy those functions or goals in particular contexts. However, different people can have different goals or purposes for the very same system.

In general, these different goals or functions will be correlated with the different roles of the people for whom the autonomous system is relevant. We use this somewhat awkward phrasing because not every such person is a "user" of the system in the traditional sense. We can usefully, though roughly, divide these individuals on the basis of their typical goals into three groups that we will call: Engineers, Users, and Affectees (i.e., those who are affected by the system). These groups are neither disjoint nor eternal: the same individual can have different goals (and so belong to multiple groups) at a time, and their goals can shift over time. In addition, one or more of these groups might be empty for a particular system at a particular time.

Engineers are (in our taxonomy) the people who have the goals of designing, developing, and deploying the autonomous system, often in a professional capacity. For example, Engineers for a face recognition system are primarily the development and implementation teams. Engineers have the goal and ability to shape the design and operations of the system so it can perform the intended functions. For them, the "inner workings" of the system are directly relevant, since success as an Engineer is based on the system appropriately performing its intended functions, whatever they might be. The autonomous system need not directly advance the other goals or interests of the Engineers themselves (unless they are also Users or Affectees).

Users are the people who engage directly with the deployed autonomous system to achieve a personal goal (often connected with their job). In some cases, they will have specified the functions for the technology; in other cases, they will be using technology developed for a different group or purpose. For example, the Users of the face recognition system would be the people who apply it to different images and receive the tags or identities of the photographed individual. Crucially, the exact underlying algorithm is almost always irrelevant to a User. They will rarely have access to the "inner workings" of the system, but also do not need such access. Users only need to know the capabilities and functions of the system (still a big demand!) so that they can use it appropriately to reach their goals. They do not necessarily need to know *how* the system achieves those ends.

Finally, Affectees are the people who do not directly interact with the autonomous system, but whose goals are affected by its performance. For example, someone who is identified by the face recognition system would be an Affectee, as her ability to reach her goals would potentially be impacted by that identification. In fact, the whole general public may fall into the Affectee group, if the system's deployment is sufficiently large-scale. In general, the goals of Affectees will typically be quite diverse. One might object that intelligibility is irrelevant for Affectees, but there are well-established moral requirements, and perhaps legal requirements, for explanations or intelligible justifications when one is affected by a decision.

Our taxonomy is fundamentally based on people's goals or intended uses of the technology, rather than their social role (in contrast with [19]). As noted above, those goals will often *correlate* with roles, but that need not be the case. For example, anyone with the goal of understanding the inner workings of the system will qualify as an Engineer, regardless of whether they actually fill that social or professional role in the development process.

Consider now the possible function of intelligibility. In general, the black box nature of an autonomous system could potentially lead to lack of trust, or even mistrust. However, the particular trust needs of an individual can differ depending on the ways in which the person engages with, or is affected by, the system [13]. Moving beyond trust, we might plausibly require different information—different types of intelligibility—depending on whether we are evaluating the moral, economic, psychological, social, or other impacts of some black-box autonomous system [6].

Most importantly, the intelligibility (or not) of a system matters only in as much as it supports or frustrates the goals and interests of relevant people. If some individual is able to accomplish *all* of their goals with a black box, unintelligible system, then it is quite unclear what reason that individual would have to demand a change in the system [9]. In particular, someone who cares *only* about accuracy will be able to accomplish all of their goals with a black-box system, and so will see no reason to care about whether a system is intelligible. In fact, intelligibility may well be an impediment to that goal, and so they are right not to care about whether a system is intelligible.

Of course, other people might have different goals, and for some of them, some degree of intelligibility may be important. The error in the simple argument is that it assumes that accuracy (or other goal-independent performance measure) is the only plausible goal, rather than recognizing the wide range of goals that people might have. And since intelligibility of an algorithm or autonomous system depends on one's goals and knowledge, we must consider the possibility that there are multiple notions of 'intelligible'. We thus turn to ask: Is there a natural "taxonomy" of intelligibility?

2 Varieties of Intelligibility

One might hope that there would be just one relevant notion of intelligibility, perhaps with some "free parameters" (e.g., background knowledge of the relevant individual). Unfortunately, the diversity of goals across potentially-affected groups implies qualitatively different notions of intelligibility. At the same time, we argue here that the limited number of sets of goals for those distinct groups prevents an unwieldy proliferation of notions. Of course, people can have many different goals at the same time, and so someone can require more than one kind of intelligibility given her specific goals.

We focus here on "normative" intelligibility that obtains independently of whether some individual descriptively happens to believe a system to be intelligible. In particular, whether a system appears intelligible to someone surely depends on their attention, past history, and so forth. We focus here on the ways in which intelligibility depends on people's goals: those goals imply corresponding reasons to need such understanding, and we can ask when a system provides such reasons.

2.1 Difference-making Intelligibility

First, consider the type of intelligibility required by the Affectees—individuals whose goals are (potentially) affected by the system's operation, but who cannot directly use or build it. These individuals cannot advance their goals by impacting the "inner workings" of the autonomous system, nor by changing the broader sociotechnical system within which it functions (without substantial effort). Rather, Affectees only have control over (at most) the inputs that are provided to the system. Hence, the intelligibility that they require is of a difference-making sort [1, 22]: if the inputs to the system had been different, then how (if at all) would the system have behaved differently?[1]

This type of intelligibility is related to, but different from, standard measures of reliability. Both depend on how the system would have performed under various changes to the inputs. On the one hand, though, Affectees need to understand (not simply measure) the ways in which different inputs could make a behavioral difference [5]. On the other hand, Affectees do not necessarily need the level of quantitative performance information required to measure the precise reliability. Sometimes, intelligibility for Affectees only requires knowing the qualitative difference-making relations.

Since Affectees do not interact directly with the system, intelligibility should aim to provide an input-output characterization of the decision processes embodied by the algorithm, thereby reducing uncertainty and demonstrating the reliability of the system. The outputs should be explained (in a difference-making way) by the contributing features. For example, in our face recognition system, an Affectee needs to know what aspects of their physical appearance, the environmental conditions, and the sensor states made a meaningful difference in the system identifying them. Said differently, what changes would have made a difference in the identification? Of course, the Affectee might not be able to change some of those factors (e.g., the sensor state), but this information is what matters for their goals.

This type of intelligibility is unusual, as the difference-making relationships need not correspond in any direct way to the actual operation of the system. As a simplistic example, one could know that the input-output relationship is $f(x, y) = x^2 - y^2$ without having any further information about which algorithm (of the infinitely many possibilities) is actually used to compute it.[2] Such information is irrelevant for the goals of the Affectee, and so they are not required for intelligibility (for them). Affectees only need to know how things would have been (or could be) different if the input varied in particular ways.

On a practical level, difference-making intelligibility can be ensured with minimal assumptions about people's prior knowledge. They need not know the specific models and algorithms used in implementation, and so technical jargon and knowledge can, and should, be avoided where possible. In many cases, more common or colloquial descriptions (and perhaps even not-necessarily-true descriptions; see below) will be entirely appropriate. This type of intelligibility is thus fairly straightforward to provide. Difference-making inputs can be identified using "test queries" of various types, and so even deep networks can be intelligible in this difference-making sense [11]. Alternately, one could perform indirect causal inference on the decision process within the autonomous system [4]. This latter process could even potentially identify some aspects of the (unobserved) internal causal processes of the system [18, 23], though those are not necessary for this type of intelligibility.

Finally, difference-making intelligibility can enhance trust in autonomous systems by eliminating perceived uncertainties and assuring reliability in the system. Because it focuses on the input features as difference-making causes of outputs (in the system), it can also help to expose algorithmic biases that are relevant to the goals of Affectees. It potentially provides a common language for discussions of reliability, fairness, and bias in terms of impacts due to particular (difference-making) inputs.

2.2 Function-based Intelligibility

While intelligibility in terms of difference-making is relatively low-cost, it also has limited benefits for Users and Engineers. Those groups have different goals that require them to know more than simply the input-output relationships. In particular, a User's goal of successfully using the system for some purpose requires information about the intended or normal functions of the system. A second type of intelligibility thus explains why those input-output relations obtain, through a specification of the "contentful" functions performed by the system. That is, this type of intelligibility describes "what the system is trying to do."[3]

Function-based intelligibility enables Users to effectively integrate the system into their particular use cases, precisely because it provides the "content" of the system's outputs and processing, as well as the contexts of relevant applicability (and non-applicability!). For example, a User of the face recognition system would need to know what information is provided by the identification, as well as the conditions under which it does (or does not) perform to acceptable levels. This understanding can

[1] Importantly, these are difference-making relations *for the autonomous system*, which need not correspond to causal difference-makers in the world. For example, a symptom could be a difference-maker for a medical diagnosis AI (i.e., a change in the input symptom would lead to a different diagnosis), even though that symptom is not a difference-maker in the world (since intervening on the symptom will not cure the disease).

[2] This type of intelligibility is related to computational-level explanations in cognitive science [10], since implementation details are irrelevant for both. Importantly, though, computational-level explanations also involve claims about the optimality of the computed function for the agent's problem. Nothing like that claim is required here.

[3] Function-based intelligibility is thus closer to computational-level explanations (see fn. 2), as it connects the system with the problem that the User is trying to solve.

lead to more appropriate and ethical use, including a decision to not use a system when the conditions not appropriate for that use.

This type of intelligibility does not necessarily describe internal causal structures or even difference-making input-output relationships, though the latter will usually be implied by the contentful functions. Rather, the focus is on the functionality of the system, including appropriate conditions for its use or adaptation. As a result, this type of intelligibility supports a richer type of trust compared to "mere" reliability [13]. Trust based solely on performance reliability can be quite fragile, since one does not necessarily know whether some new case falls within the normal range. In contrast, this type of intelligibility enables the User to know the contexts in which she can accomplish her intended goals using the system.

Function-based intelligibility will typically be straightforward to realize if developers are involved in the process. Much of the relevant information can be found in design documents for the system, as those should express the intended functions and contexts of use for the system.[4] Of course, these documents will not necessarily be helpful if the developers failed to fully test, validate, and verify their system. However, they provide a way to specify the sometimes-vague functions of a system. Those documents and related developer knowledge may require some translation to be expressed in ways that are understandable by the intended Users. However, since those Users will typically be known in advance, this requirement is relatively minimal.

In general, the low cost of this path is quite valuable given the significant potential benefits of this type of intelligibility. It provides Users with exactly the abilities that they require to efficiently integrate, appropriately use, and accurately interpret the performance of the autonomous system so that they can reach their goals. This type of intelligibility does not, however, answer all possible why-questions, particularly those that focus on the internal operations of the system.

2.3 Causal-process Intelligibility

Engineers require more understanding of a technology than is provided by either of the previous two notions. In particular, the goals of Engineers can require them to intervene and change the inner workings of an autonomous system to improve performance. This level of understanding involves the ability to provide causal process explanations [16] of the system behavior. This type of strong intelligibility is arguably the sole focus of many current discussions of intelligibility for autonomous systems, but we emphasize that it is only one of several types, and it is only required for those with the goals of Engineers.

These explanations will typically involve information about the overall computational architecture, specific models, parameter values, internal states, and their relationships (or at least, a significant subset of these).[5] There may also be hardware or user interface constraints that do not directly impact the system's underlying causal processes, but do provide valuable information about other design decisions (and perhaps also constraints on the changes or interventions that an Engineer could perform). For example, Engineer intelligibility for the face recognition system would require knowledge of the particular learning algorithms, training data and methods, post hoc adjustments, and all of the other elements that went into the development and deployment of the system.

Although there are heavy content requirements for this type of intelligibility, one can also make substantial assumptions about the knowledge of the intended population of Engineers. These individuals will typically be highly educated in relevant frameworks, concepts, and jargon, and so the efforts for intelligibility can focus on the content. In addition, this shared background knowledge means that the space of relevant possible causal processes is further constrained relative to, say, the beliefs that Affectees might have about how these systems work. Implementation of this type of intelligibility can focus on the processes that are distinctive for this particular autonomous system.

At the same time, we should not minimize the difficulty of specifying the relevant aspects of the internal causal processes. For example, deep networks are almost never intelligible in this way (at least, given existing methods), precisely because the internal logic and causal processes are too complex to be understood. In other cases, the challenging complexity of the system arises from interactions between various system components. In some systems, each individual subsystem is itself intelligible, but the subsystems influence each other in complex, non-intelligible ways. Causal-process intelligibility may simply not be possible for some systems or architectures, given current methods to extract and represent internal causal processes.

One significant barrier to causal-process intelligibility is the possibility that key information must be kept confidential. In many cases, the most difficult part of system development is not identification of appropriate functions or goals for the system, but rather the fine-grained details about how to succeed at those functions. Hence, companies will often be quite reluctant to reveal the information required for this type of intelligibility. However, notice that this information is only required by Engineers. The other two types of intelligibility are available without revealing proprietary knowledge or information.

When causal-process intelligibility is available, it can provide significant benefits. First, Engineers can adjust or change the system in a targeted manner to improve performance or widen the contexts of applicability. They can thereby better reach their goals of producing a better-functioning (according to some specification) system. These focused interventions cannot be effectively

[4] Of course, this assumes that developers document their design thinking.

[5] This type of intelligibility thus closely corresponds to algorithmic- or implementation-level explanations in cognitive science [10].

performed given only the previous two types of intelligibility, so individuals with improvement goals require this stronger notion. Second, causal-process intelligibility can enable the Engineer to adapt the system to novel functions or uses. Knowledge of the underlying causal processes enables modifications that alter the system's functions, and so can expand those functions or their contexts of use.

3 Implementing Intelligibility

The previous section outlined three different notions of intelligibility, each of which provides the type of information that is required to achieve some set of goals through the technology. Intelligibility is not a "one size fits all" property, but rather imposes different requirements depending on whether one has the goals of a User, Engineer, or Affectee. We now turn to issues that are both challenges and opportunities for our account. In particular, (a) the world of goals is more complicated than our simple characterization; (b) intelligibility as we have framed it is not readily assessed; and perhaps most worrisome (c) our account implies that false statements about an autonomous system can sometimes increase its intelligibility. We address each of these in turn.

3.1 Diversity of Goals, Diversity of Groups

First, the space of possible goals is obviously much richer than our framework suggests. People's goals do not perfectly cluster into those of User, Engineer, and Affectee. Rather, an individual's interests and goals can (i) shift over time; (ii) involve goals from multiple groups; and (iii) fall outside of the main goals that we considered.

Possibility (i) implies that people can belong to different groups over time, and so require different notions of intelligibility at different times. For example, a member of the general public (an Affectee of the face recognition system) might join a relevant government agency (and so acquire the goals of a User). That same individual might later start to modify and adapt the system (so have an Engineer's goals). At each moment in time, though, there is no ambiguity or conflict in the information that they require.

Possibility (ii) implies that someone could simultaneously be a member of multiple groups (e.g., she could simultaneously have the goals of using and adapting the system). If the information required for different types of intelligibility are incompatible with one another, then it will be impossible to provide this individual with all of the intelligibility that she requires for success at her goals. We might be forced to choose, for these individuals, between different types of intelligibility for some system.

We agree this is a potential worry, but we suspect that it will rarely present a significant practical challenge, as the information for the varieties of intelligibility are related (albeit, not hierarchically). At a high level, moving from intelligibility for Affectees to Users to Engineers largely involves providing increasingly more information.[6] For example, Engineers often need information about counterfactual performance or difference-making relationships [20], and that is exactly what someone with Affectee goals requires. Moreover, we conjecture that people rarely actually have the goals of multiple groups at a specific moment in time. In practice, people seem to be more likely to rapidly cycle between sets of goals, rather than have cross-group goals at a single instance.

Possibility (iii) implies that our taxonomy of intelligibility is incomplete since we do not cover all possible goals. We acknowledge that possibility, though we also note that our characterizations of the goals of the three groups were quite broad (and many other goals will be irrelevant to intelligibility). More importantly, we do not here claim completeness; if a different group were to be identified with different core goals and interests, then this framework can be extended to include a notion of intelligibility that provides them with the information and understanding that they require to achieve their goals through the technology.

3.2 Evaluating Intelligibility

Second, we have not seriously considered questions of evaluation: how do we determine whether a system is actually intelligible for people with certain goals? Systematic, consistent evaluation not only enables iterative improvement, but also clarifies the affected group's needs and requirements. Current efforts to measure intelligibility often rely on introspective self-reports from people, usually through responses to explicit questions about explanations, intelligibility, or understanding [11]. However, these measures address only "descriptive" intelligibility, not the normative target of our analysis. Thankfully, there are other measures that can be used to assess whether people have appropriate reasons to understand how a system does (or does not) support their goals.

In general, an individual's understanding of an autonomous system could be implicit, just like much of our understanding of the world [12]. For example, repeated experiences could enable a User to employ an autonomous decision-maker only in appropriate contexts, but without any conscious, explicit representation of those contexts. In such a case, the system is (on our framework) intelligible for that User, since they have the necessary (implicit) understanding of the system to effectively advance their goals. Hence, our evaluations must be sufficiently general to detect and measure this type of "intelligibility through repeated experiences" (and other indirect ways of producing intelligibility for a group).

At the same time, we expect that 'intelligibility' will principally be a useful guide for developers and deployers who want to explicitly teach people with particular goals. Implicit learning is an important part of our cognition, but it also can be significantly

[6] The exception is that contentful functions do not completely subsume difference-making relationships. In ordinary circumstances, though, the former will include or imply almost all of the latter information.

slower than explicit instruction [14, 17]. Hence, we expect that it will be more valuable to have explicit measures of (normative) intelligibility if one is trying to ensure that people have appropriate understanding.

Finally, our measures must also be sensitive to potential differences in goals between various subgroups. For example, the User goals might actually separate into Operator and Executor goals (parallel to the role-based subgroups described in [19]). We leave those finer-grained distinctions to future research.

3.3 Intelligibility through Falsehoods

Third, a surprising implication of our framework is that false claims can actually increase intelligibility (sometimes). This implication might appear to be a fatal flaw: intelligibility seems to be closely tied to knowledge, and so one might object that falsehoods should never increase intelligibility. Instead, we contend that this surprising implication is a feature, not a bug of our framework.

Our notion of intelligibility emphasizes its role in changing people's understanding of an autonomous system so that they are better enable them to reach their goals and realize their (relevant) interests. And it turns out, perhaps surprisingly, that false beliefs can actually improve our ability to reach our goals, as they can help to make correct action more likely [21]. For example, a slightly inaccurate characterization of the exact function of the face recognition system need not impair a User's ability to achieve her goals, if the incorrect implications and inferences from a false belief pertain only to conditions the User never encounters. Moreover, these false beliefs might even help her to achieve her goals, if they provide other advantages (e.g., the false beliefs better cohere with expectations).

In fact, there are numerous real-world examples of this phenomenon: namely, when falsehoods enter through simplifications that omit irrelevant details. The resulting information and beliefs are not strictly true, but those inaccuracies do not impair an individual's ability to reach her goals. And in light of the many advantages to simple-and-broad explanations and theories [8], those inaccuracies can actually improve her ability to achieve her goals with respect to the system. We conclude that a focus on true information and true beliefs is an appropriate default position for any implementer of intelligibility, but they should recognize that the key question is how the information impacts an individual's success at reaching their goals, not whether it leads to strictly true beliefs.

4 Conclusions

We have here focused on a pragmatic understanding of intelligibility as a property that advances people's interests and abilities to reach their goals. Intelligibility is thus not an intrinsic property of a system, but rather requires careful consideration of the people who will use, develop, or be affected by the system. As we have argued here, the result of this alternative perspective is a taxonomy of different notions of 'intelligibility'—difference-making, function-based, and causal-process—that require different types of information. Intelligibility is not a one-size-fits-all matter, but rather must be tuned to the particular needs of the people who seek to understand the system (for their goals).

ACKNOWLEDGMENTS

This paper was made possible in part by support from an Andrew Carnegie Fellowship (Carnegie Corporation of New York) to DD. The statements made and views expressed are solely the responsibility of the authors.

REFERENCES

[1] S Beckers and J Vennekens, 2018. A Principled Approach to Defining Actual Causation. *Synthese* 195: 835-862.

[2] D Doran, S Shulz and T R. Besold, 2017. What Does Explainable AI Really Mean? A New Conceptualization of Perspectives. In T. R. Besold & O. Kutz (Eds), *Proceedings of the First International Workshop on Comprehensibility and Explanation in AI and ML 2017.*

[3] D Gunning, 2017. Explainable Artificial Intelligence (xai). *Defense Advanced Research Projects Agency (DARPA).*

[4] M Harradon, J Druce and B Ruttenberg, 2018. Causal Learning and Explanation of Deep Neural Networks via Autoencoded Activations. arXiv:1802.00541.

[5] P Lipton, 2003. *Inference to the Best Explanation.* Routledge.

[6] Z C. Lipton, 2016. The Mythos of Model Interpretability. arXiv:1606.03490.

[7] Z C. Lipton, A Chouldechova and J McAuley, 2018. Does Mitigating ML's Impact Disparity Require Treatment Disparity? arXiv:1711.07076.

[8] Tania Lombrozo, 2016. Explanatory Preferences Shape Learning and Inference. *Trends in cognitive sciences* 20(10): 748-759.

[9] Alex J. London, 2019. Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability. *Hastings Center Report* 49: 15-21.

[10] David Marr, 1982. *Vision.* San Francisco: W.H. Freeman.

[11] G Montavon, W Samek and K R. Müller, 2018. Methods for Interpreting and Understanding Deep Neural Networks. *Digital Signal Processing* 73, 1-15.

[12] A S. Reber, 1989. Implicit Learning and Tacit Knowledge. *Journal of Experimental Psychology: General* 118: 219-235.

[13] H M. Roff and David Danks, 2018. "Trust but Verify": The Difficulty of Trusting Autonomous Weapon Systems. *Journal of Millitary Ethics* 17: 2-20.

[14] D A. Rosenbaum, R A. Carlson and R O. Gilmore, 2001. Acquisition of Intellectual and Perceptual-motor Skills. *Annual Review of Psychology* 52: 453-470.

[15] C Rudin, 2018. Please Stop Explaining Black Box Models for High Stakes Decisions. *NeurIPS 2018 Workshop on Critiquing and Correcting Trends in Machine Learning.* arXiv: 1811.10154v2

[16] W C. Salmon, 1984. *Scientific Explanation and the Causal Structure of the World.* Princeton: Princeton University Press.

[17] D R. Shanks, 2010. Learning: From Association to Cognition. *Annual Review of Psychology* 61: 273-301.

[18] R Silva, R Scheines, Clark Glymour and P Spirtes, 2006. Learning the Structure of Linear Latent Variable Models. *Journal of Machine Learning Research* 7: 191-246.

[19] R Tomsett, D Braines, D Harbone, A Preece and S Chakraborty, 2018. Interpretable to Whom? A Role-based Model for Analyzing Interpretable Machine Learning Systems. arXiv:1806.07552.

[20] D S. Weld and G Bansal, 2018. The Challenge of Crafting Intelligible Intelligence. *Communications of the ACM* 62(6), 70-79.

[21] S Wellen and David Danks, 2016. Adaptively Rational Learning. *Minds & Machines* 26: 87-102.

[22] J Woodward, 2005. *Making Things Happen: A Theory of Causal Explanation.* Oxford: Oxford University Press.

[23] X Zhang, K Korb, A Nicholson and S Mascaro, 2017. Applying Dependency Patterns in Causal Discovery of Latent Variable Models. *Lecture Notes in Computer Science Vol 10142* (pp. 134-143). Springer Verlag.