Challenges for Transparency

Adrian Weller [1 2 3]

Abstract

Transparency is often deemed critical to enable effective real-world deployment of intelligent systems. Yet the motivations for and benefits of different types of transparency can vary significantly depending on context, and objective measurement criteria are difficult to identify. We provide a brief survey, suggesting challenges and related concerns. We highlight and review settings where transparency may cause harm, discussing connections across privacy, multi-agent game theory, economics, fairness and trust.

1. Introduction

The case for transparency has been made in many settings, including for government policy (Vishwanath and Kaufmann, 2001), business (Lowenstein, 1996), charity (Sridhar and Batniji, 2008), and algorithms (Mortier et al., 2014). Within machine learning, there is a general feeling that "transparency" – like "fairness" – is important and good. Yet both concepts are somewhat ambiguous, and can mean different things to different people in different contexts. We discuss various types of transparency in the context of human interpretation of algorithms, noting their benefits, motivations and difficulties for measurement.

We then consider settings where, perhaps surprisingly, transparency may lead to a worse outcome. Transparency is often beneficial but it is not a universal good. We draw attention to work in other disciplines and hope to contribute to an exploration of which types of transparency are helpful to whom in which contexts, while recognizing when conditions may arise such that transparency could be unhelpful.

We summarize our main themes:
(A) There are many types of transparency with different motivations – we need better ways to measure them (§2).
(B) We should recognize that sometimes transparency is a means to an end, not a goal in itself (§2 and §3.3).
(C) Actors with misaligned interests can abuse transparency as a manipulation channel, or inappropriately use information gained (§3).
(D) In some settings, more transparency can lead to less efficiency (§4 reviews economics, multi-agent game theory and network routing), fairness (§5) and trust (§3.2 and §6).

In §7, we raise 'machine interpretability' as an important research direction, which may also provide insight into how to measure human understanding in some settings.

Related Work There is a considerable literature on transparency and social good. Much of this focuses on the benefits of transparency but some earlier work, notably in economics and social science, also considers drawbacks of transparency and accountability (Lerner and Tetlock, 1999; Prat, 2005; Etzioni, 2010; Peled, 2011; Bernstein, 2014). We discuss related work throughout the text.

2. Types of Transparency: Benefits, Measurement and Motivations

In this section, we briefly describe various types of transparency in the context of human interpretability of algorithmic systems, highlighting different possible motivations. We typically seek an explanation in understandable terms, which can often be framed as answering questions of "what", "how", or "why" (either *toward what purpose* in the future, or *due to what cause* in the past). Some of our observations have been made previously (Lipton, 2016; Doshi-Velez and Kim, 2017; McAllister et al., 2017). To our knowledge, in the machine learning setting, we make new points on motivations and on measuring understanding.

An automated explanation might arise immediately from the original system – typically if it has been constrained to lie in some set of classifiers deemed to be interpretable (e.g. a short decision list). Alternatively, a second explainer algorithm may have produced an explanation for the original system. We consider various classes of people: a *developer* is building the system; a *deployer* owns it and releases it to the public or some user group; a *user* is a typical user of the system. For example, developers might be hired to build a personalized recommendation system to buy products, which Amazon then deploys, to be used by a typical member

[1]Department of Engineering, University of Cambridge [2]Alan Turing Institute [3]Leverhulme Centre for the Future of Intelligence. Correspondence to: Adrian Weller <first.last@eng.cam.ac.uk>.

of the public. People might be experts or not.

We list several types and goals of transparency. Each may require a different sort of explanation, requiring different measures of efficacy:

1. For a developer, to understand how their system is working, aiming to debug or improve it: to see what is working well or badly, and get a sense for why.

2. For a user, to provide a sense for what the system is doing and why, to enable prediction of what it might do in unforeseen circumstances and build a sense of trust in the technology.

3. For society broadly to understand and become comfortable with the strengths and limitations of the system, overcoming a reasonable fear of the unknown.

4. For a user to understand why one particular prediction or decision was reached, to allow a check that the system worked appropriately and to enable meaningful challenge (e.g. credit approval or criminal sentencing).

5. To provide an expert (perhaps a regulator) the ability to audit a prediction or decision trail in detail, particularly if something goes wrong (e.g. a crash by an autonomous car). This may require storing key data streams and tracing through each logical step, and will facilitate assignment of accountability and legal liability.

6. To facilitate monitoring and testing for safety standards.

7. To make a user (the audience) feel comfortable with a prediction or decision so that they keep using the system. Beneficiary: deployer.

8. To lead a user (the audience) into some action or behavior – e.g. Amazon might recommend a product, providing an explanation in order that you will then click through to make a purchase. Beneficiary: deployer.

We can differentiate between the intended *audience* of an explanation and the likely *beneficiary* (or beneficiaries). We suggest that types 1-6 are broadly beneficial for society provided that explanations given are *faithful*, in the sense that they accurately convey a true understanding without hiding important details. This notion of faithful can be hard to characterize precisely. It is similar in spirit to the instructions sometimes given in courts to tell "the truth, the whole truth, and nothing but the truth".

Defining criteria and tests for practical faithfulness are important open problems. We suggest that helpful progress may be made in future by focusing on one particular context at a time. We make a similar suggestion for the challenges of characterizing if an explanation is good at conveying faithful information in understandable form, and if a human has

Figure 1. An image (left) is given to a classification system. A separate explaining algorithm extracts the sub-image (right) which it estimates led the original system to output "beagle" (from Dabkowski and Gal, 2017). Would it be better or not if the sub-image contained the legs of the dog? That depends on the classification system and is not simple to answer, demonstrating one challenge in defining a quantitative measure of the quality of the explanation.

actually understood it well.[1] Doshi-Velez and Kim (2017) suggest several methods, such as establishing a quantitative approximate measure (e.g. if we are restricting models to be decision trees then we can feel reasonably confident that a model becomes harder to understand as the number of nodes increases), or asking a human if they can correctly estimate what the system would output for given inputs. We suggest further approaches below and highlight a key challenge.

We use (as in Doshi-Velez and Kim, 2017) the terms *global* interpretability or explanation for a general understanding of how an overall system works, as in our transparency types 2-3; and *local* interpretability for an explanation of a particular prediction or decision, as in types 4, 5, 7 and 8 (though both forms may be useful for a given type).

For global interpretability, we mention two interesting possible approaches due to quotes attributed to the physicist Richard Feynman: (i) "What I cannot create, I do not understand" suggests that in some settings, a good test of understanding might be to see if the person could recreate the whole system (given expert help and allowing some reasonable tolerance); (ii) "If you can't explain it to a six year old, you don't really understand it" suggests a possible meta-approach to test clarity of an explanation – for any given test T of human understanding, ask the person to explain the system to someone new, then give that new person the test T. Lakkaraju et al. (2016) introduced measures of human interpretability based on being able to describe a decision boundary, which would facilitate model reconstruction. We discuss further notions of understanding in §7.

Highlighting a key challenge even for local interpretability, consider explaining the output of an image classification system, as illustrated in Figure 1. Recent approaches attempt to identify the parts of a given image which are most salient,

---

[1]Greater faithfulness of an explanation may challenge the ability of its audience to understand it well, perhaps requiring a greater investment of time and effort (Ribeiro et al., 2016).

i.e. those parts which in a sense were most responsible for leading to the system's prediction (Ribeiro et al., 2016; Zintgraf et al., 2017). Such approaches can be very helpful – e.g. for type 1 transparency, we might learn that a system which reliably tells apart wolves from huskies on a test set might in fact be relying on the presence or absence of snow in the background, rather than features of the animal itself, and hence is unlikely to generalize well on test data (Ribeiro et al., 2016). Suppose we are given two such methods of generating a salient sub-image for a classification system. How should we measure which method provides a better explanation of what the system is doing? This is an important open question. Note that it is not helpful to compare against what a human thinks is relevant in the image. Rather, we want a sub-image which is high in predictive information *for the system* yet is not too large, focusing only on the relevant region. A promising possible direction for a quantitative solution was suggested by Dabkowski and Gal (2017), who propose a measure of concentration of information.

Transparency as a Proxy Transparency can provide insight into other characteristics which may be hard to measure (as noted by Doshi-Velez and Kim, 2017). In the previous paragraph, we saw how local explanations for an image classification system – revealing how wolves were differentiated from huskies – demonstrated the lack of robustness of the system. Other features where transparency can provide helpful insight include fairness, verification and causality.

3. Possible Dangers of Transparency

In this section, we begin to examine ways that transparency may be unhelpful.

3.1. Divergence between Audience and Beneficiary

There are some forms of transparency, such as types 7 and 8 in §2, where the intended audience for an explanation diverges from the beneficiary, hence the motivation may be suspect. This can lead to worrying types of manipulation and requires careful consideration.

Considering type 7 transparency, we draw attention to the remarkable 'Copy Machine' study. Langer et al. (1978) arranged for researchers to try to jump in line to make a few photocopies at a busy library copy machine. The researcher either (i) gave no explanation, asking simply "May I use the xerox machine?"; (ii) provided an 'empty' explanation: "May I use the xerox machine, because I have to make copies?"; or (iii) provided a 'real' explanation: "May I use the xerox machine, because I'm in a rush?" The respective success rates were: (i) 60%; (ii) 93%; and (iii) 94%. The startling conclusion was that saying "*because* something" seemed to work very effectively to attain compliance, even if the 'something' had zero information content. Hence, a

possible worry is that a deployer might provide an empty explanation as a psychological tool to soothe users.

In fact there is a line of research which considers all communication often to be more a form of manipulation than a way to transfer information. This view is prominent when taking an evolutionary view of multiple agents (Wiley, 1983), a perspective which we revisit throughout the remainder of this paper. Earlier work by Adelberg (1979) explored whether disclosures provided in financial reports are more aptly described as communication or manipulation.

For type 8 transparency, where the deployer has a clear motive which may not be in the best interests of the audience of the explanation, particular care and future study is warranted. Even if a faithful explanation is given, it may have been carefully selected from a large set of possible faithful explanations in order to serve the deployer's goals.

3.2. Government Use of Algorithms

In many states in the US, a private company provides the COMPAS system to judges to help predict the recidivism risk of a prisoner, i.e. the chance that the prisoner will commit a crime again if released. This information is an important factor in parole hearings to determine whether to release prisoners early or to keep them locked behind bars. Significant attention has focused on whether or not the prediction system is fair (Angwin et al., 2016). We discuss connections between fairness and transparency in §5. Here we consider the appropriate degree of transparency of such a system: a prisoner should at least have transparency type 4 from §2, but can there be too much transparency?

Perhaps motivated by concerns in the US over the COMPAS system, Bulgaria recently passed legislation requiring that (many forms of) government software be open source, "after all, it's paid by tax-payers' money and they should... be able to see it" (Coldewey, 2016). Note that a machine learning system typically consists of both algorithms and data, and having access to just one of these may not provide much meaningful information. We consider the case of all algorithms and data being transparently available and note several concerns.

Gaming and IP Incentives If all details are readily available, this can facilitate gaming of the rules (Ghani, 2016). In addition, if all code and data is open source, then this removes incentives to develop relevant private intellectual property, which may delay progress significantly.

Transparency and Privacy Indeed, in many cases, transparency may be viewed as the opposite of privacy. Many in society feel that some sort of right to privacy – and hence, a limit to transparency – is appropriate. This can create tensions even for one user whose data is used in a system

– the user may want their personal data to be kept private but might also like a right to an explanation of how that same system (algorithms + data) works. Further, there are many settings where privacy (i.e. a lack of transparency to all) is critical to foster a trusting relationship of confidence. Examples include the relationship between a doctor and a patient, a lawyer and their client, or discussions of international diplomacy. Inside these relationships, it is interesting to question whether greater transparency leads to trust. We return to this topic in §6. Here we rather suggest that a prudent approach is to release private information only to a partner that is already trusted, hence trust can lead to transparency. Providing information to an agent empowers them (Solove, 2001), hence you should first be confident that their interests align with yours.

As one example, there has been discussion about the extent to which government agencies such as the NSA should be allowed to collect data on individuals. Some argue that "if you've got nothing to hide, you've got nothing to fear" (Solove, 2011) or suggest that collecting only 'metadata' is harmless. Yet Cole (2014) quotes General Michael Hayden, former director of the NSA and CIA, as saying "We kill people based on metadata."

3.3. Means and Goals

In §2, we noted that transparency can serve as an imperfect proxy to gain insight into other desirable properties of a system, such as reliability or fairness. For example, transparency is often cited as critical for deployment of autonomous vehicles. We suggest that these transparency concerns are primarily for types 1, 2, 3, 5 and 6. Each of these involves somewhat different types of explanation or understanding, with some easier to implement than others. A key concern is reliable safety: how can we be certain that the vehicle will perform well in all circumstances if we do not understand exactly how it is working?

In such cases, we should take care not to stifle innovation by confusing transparency as an end in itself rather than a means to other goals. It is conceivable that much time and resources could be spent trying to gain an extremely transparent understanding, when those efforts might be better spent directly on the goal of improving safety. Being pragmatic, it is plausible that society will resort to implementing various safety tests, such as are used for aircraft autopilot systems. If such tests are passed by an autonomous vehicle system, and accidents are extremely rare in practice, then it may make sense to proceed even without full transparency. After all, which is preferable: full transparency and many deaths per year from accidents, or much less transparency and far fewer deaths per year? As Einstein (1942) eloquently put it, we must beware the "perfection of means and confusion of goals".

Initial configuration. Everyone has cost (delay) of 83.
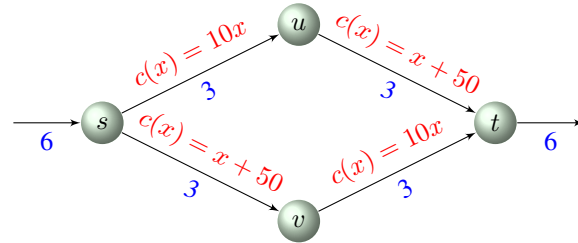
Edge $u \to v$ is added. Everyone has cost (delay) of 92.
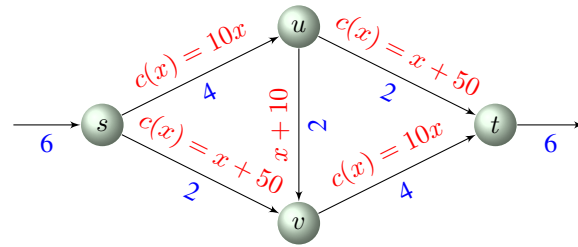

Figure 2. An example of Braess' paradox for network flow from (Kelly, 1991). 6 cars of flow must pass from s to t. For each edge: red (above) shows the cost (i.e. delay incurred) as a function of the flow x through it; blue (below) shows the Wardrop equilibrium flow given the costs, based on each car selfishly optimizing. When the edge $u \to v$ is added, surprisingly everyone does worse.

4. Economics & Multi-agent Game Theory

In an economy, each individual acts as an autonomous agent. If each agent optimizes her own selfish utility, there is no guarantee in general that this will lead to the best outcome for society. Under restrictive assumptions, Arrow and Debreu (1954) famously proved the existence of a general equilibrium for a competitive economy. If *externalities* are present, i.e. if costs of one agent's actions fall on others, then we should expect that the result of each agent optimizing her own outcome may lead to a suboptimal result for the whole. This phenomenon is sometimes described as the *price of anarchy* (Roughgarden, 2006).

We relate this to transparency by considering what happens if all agents are given additional faithful information about a system. This kind of transparency is closest to type 3 from §2, which may seem the most innocuous of all. An engineering perspective might naturally lead one to suspect that more information should lead to a better outcome. However, the background of an economist or multi-agent game theorist helps to realize that more information empowers the agents to optimize their own agendas more efficiently, and thus may lead to a worse global outcome. We illustrate these ideas with a striking example of Braess' paradox (Braess, 1968) as given by Kelly (1991).

Figure 2 (top) shows a traffic network where 6 cars enter s at the left, flow through the network via either u or v and exit from t at the right. The costs (delays) of each edge are

shown in red, and importantly rise as the amount of flow through them increases. This is realistic in that greater traffic flow on a road often leads to longer delays for everyone on it. Thus we have externalities. If each agent optimizes her own utility, the *Wardrop equilibrium* shown in blue is reached, where each car incurs a delay of $10 \cdot 3 + 3 + 50 = 83$ time units (details in Kelly, 1991; at a Wardrop equilibrium, no individual can reduce her path cost by switching routes, hence all routes have the same cost).

Now consider what happens if all cars learn about an extra road from u to v as shown in Figure 2 (bottom). We may assume that the $u \to v$ road was always there but that it was hidden until the faithful information about its existence was made transparent. Although intuitively the additional road increases capacity for all and hence should only lead to a better outcome, in fact the new Wardrop equilibrium obtained by selfish optimization results in a greater delay for everyone of $10 \cdot 4 + 2 + 50 = 92$ units!

Kelly (1991) provides an additional twist on this example. Suppose now that all users know that road works might be under way on the $u \to v$ road, which has delay $x + 10 + R$, where R is a random variable taking the values 0 or 30 with equal probability. If R is unknown, then the expected delay along $u \to v$ is $x + 25$, leading to the outcome in Figure 2 (top) with delay of 83. On the other hand, if instead R is known to everyone, then: we either have $R = 30$ which leads to the outcome in Figure 2 (top); or we have $R = 0$ which leads to the outcome in Figure 2 (bottom); for an expected delay of $\frac{1}{2}(83 + 92) = 87.5 > 83$. Hence, again transparency (providing everyone with faithful information) leads to a worse expected outcome for everyone.

5. Fairness and Discrimination

Much work on fairness in machine learning has focused on attempting to avoid discrimination against sub-groups as identified by sensitive features, such as race or gender. Typical metrics for discrimination are based on various types of *disparate impact* or *disparate treatment* (Barocas and Selbst, 2016). Here we consider a theme which relates to a common fairness approach used to avoid disparate treatment: simply remove the sensitive feature(s) from the data. A valid objection to this method is that it may be possible to predict (and hence reconstruct) the sensitive feature(s) from other features with high confidence. Nevertheless, the approach is in widespread use and the example below may help to explain its intuitive appeal to some.

Axelrod (2006) describes fascinating work exploring the conditions under which cooperation naturally emerges in multi-agent populations. In many settings, repeated contact leading to iterated prisoner's dilemma interactions are supportive of the emergence of cooperation (for details of
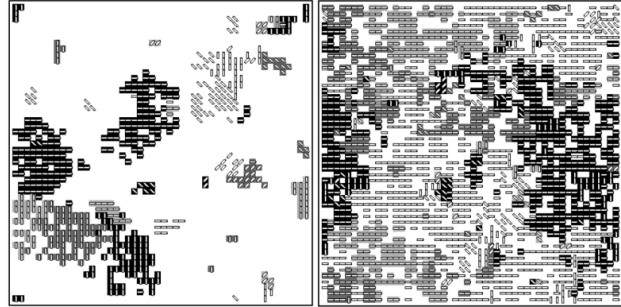

Figure 3. Illustrations by Hammond and Axelrod (2006) of the evolution of ethnocentrism. 4 different tribes are shown as shades of gray. Agents which are {Ethnocentric (cooperate with their own tribe, defect against other tribes) / Pure cooperators (always cooperate) / Egoists (always defect) / Those who cooperate only with other tribes} have {Horizontal / Vertical / Diagonal up-right / Diagonal up-left} lines respectively. The left image shows one run after 100 periods; the right image is after 2000 periods. Over time, almost all agents are ethnocentric. See the reference for a video.

the prisoner's dilemma, see `https://en.wikipedia.org/wiki/Prisoners_dilemma`).

In subsequent work, Hammond and Axelrod (2006) expanded the framework to consider agents which have an extra feature of 'ethnicity' which might be regarded as their color or tribe. Now each agent could either cooperate or defect when interacting with other agents of the same tribe; and similarly could either cooperate or defect with agents of other tribes. Using a simple model of multi-agent evolution, it was strikingly demonstrated that a robust conclusion across a wide range of model parameters, is that 'ethnocentrism' emerges. That is, after many iterations, each agent is very likely to cooperate with others of the same tribe, but to defect against agents from other tribes. See Figure 3.

Hence, in this setting, greater transparency (i.e. making faithful information on ethnicity available) leads to the emergence of discriminatory behavior.

Recent work explores multi-agent reinforcement learning in repeated social interactions to begin to identify conditions on the environment and the agents' cognitive capacities which lead to the emergence of cooperation (see Leibo et al., 2017; Lowe et al., 2017 and related blog entries).

6. Trust: Transparency and Honesty

Many support the view that transparency builds trust. For example, the Dalai Lama is reported to have warned "A lack of transparency results in distrust and a deep sense of insecurity." However, we suggest that the story is more nuanced. If we distinguish between transparency – i.e. the provision of information – and honesty – i.e. the accuracy of the information – then which is more important?

A common view is that trust relies on honesty. The following quote is attributed to Simon Sinek: "Trust is built on telling the truth, not telling people what they want to hear." However, as we can sense from the following phrases themselves, it can be difficult to hear the 'harsh truth' from someone speaking with 'brutal honesty'.

In recent work, Levine and Schweitzer (2015) demonstrated settings where *prosocial lies* increased trust. Prosocial lies are a form of deception – the transmission of information that intentionally misleads others – which benefits the target. Their work aims to separate the roles of benevolence and integrity in building interpersonal trust. They conclude that altruistic lies increase trust when deception is directly experienced, or even when it is merely observed.

These effects may make sense when we consider that humans evolved in a tribal multi-agent society. Trusting someone may reflect a belief that they are 'in our tribe' and will look out for our interests. From a historic survival perspective, this benevolence may have been more important than truthful communication – perhaps a 'pre-truth' society?

Interestingly, work in psychology indicates that we are not good at estimating how transparent we our ourselves when communicating with others. Gilovich et al. (1998) showed evidence for the *illusion of transparency* – a tendency for people to overestimate the extent to which others can discern their internal states. This is attributed to a tendency for people to adjust insufficiently from the 'anchor' of their own experience when attempting to adopt another's perspective. Consistent with this, Griffin and Ross (1991) observed the following result. If someone is asked to tap a well-known melody on a tabletop and then to estimate the chance that a listener will be able to identify the song they have tapped, tappers grossly overestimate the listener's abilities.

7. Machine Interpretability & Understanding

The focus of this workshop is on human interpretability, that is helping humans to understand machines. But two classes of 'machine interpretability', that is helping machines to understand, are also important lines for future research.

First, it will be increasingly helpful for machines to be able to understand humans and our motivations. As examples, consider automated care for the elderly, or how an autonomous vehicle waiting at a crossroad should perceive and respond reliably if a human in a car opposite is beckoning to advance.

Second, we believe that a fruitful line of work will be to help machines understand each other. Exciting work has begun to explore this direction, looking for ways to enable multiple agents to cooperate effectively (Evtimova et al., 2017; Mordatch and Abbeel, 2017). Different paradigms of

multi-agent organization should be explored (Horling and Lesser, 2004). One motivation is the goal of AI agents which can autonomously generate flexible, hierarchical *concepts* which can apply broadly, going beyond traditional transfer learning. Ideally, these concepts will capture high level structure in a way which can be communicated efficiently and deployed flexibly. Although it is highly desirable for humans to understand this structure, we suggest that it will still be extremely useful, and perhaps easier, to begin by working on ways for machines to communicate with each other. Further, complex structures may be developed which are beyond easy human understanding.

It is possible that a relatively low capacity information bottleneck (as examined by Tishby et al., 1999) will be a useful constraint between agents to develop such structures. Indeed it has been suggested that the low inter-human bandwidth of speech compared to the higher processing power of our brains may have helped lead to the development of our own intelligence and consciousness (Lawrence, 2017).

With this goal in mind, a useful metric of successful 'transfer of understanding' (in this sense, a metric for a good explanation) could be to measure how agent B's performance on some task, or range of tasks, improves after receiving some limited information from agent A – though important details will need to be determined, including suitable bandwidth constraints, and recognition that what is useful will depend on the knowledge that A has already.

A similar notion might be useful for measuring human interpretability. An example of a helpful conceptual explanation to improve a beginner's ability to play chess might be "try to control the center". One advantage is that only performance improvement need be measured, bypassing the difficult task of quantifying internal understanding directly.

8. Conclusion

There are many settings where transparency is helpful. We have described some of these settings and tried to begin to clarify just what sort of transparency may be desirable for each, with accompanying research challenges. This topic is timely given keen interest in laws such as the GDPR (enforceable in the EU from 2018) which seek to provide users with some sort of right to an explanation of algorithmic decisions.

One focus of this work is to highlight scenarios where transparency may cause harm. We have provided examples where greater transparency can lead to worse outcomes and less fairness. We hope to continue to develop frameworks to understand what sorts of transparency are helpful or harmful to whom in particular contexts. This is a rich area which can draw on connections across economics, multi-agent game theory, law, policy and cognitive science.

Acknowledgements

References

A. Adelberg. Narrative disclosures contained in financial reports: means of communication or manipulation? *Accounting and Business Research*, 9(35):179–190, 1979.

J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica, May*, 2016.

K. Arrow and G. Debreu. Existence of an equilibrium for a competitive economy. *Econometrica: Journal of the Econometric Society*, pages 265–290, 1954.

R. Axelrod. *The evolution of cooperation*. Basic books, 2006.

S. Barocas and A. Selbst. Big Data's Disparate Impact. *California Law Review*, 104(3):671–732, 2016.

E. Bernstein. The transparency trap. *Harvard Business Review*, 92 (10):58–66, 2014.

D. Braess. Über ein paradoxon aus der verkehrsplanung. *Mathematical Methods of Operations Research*, 12(1):258–268, 1968.

D. Coldewey. Bulgaria now requires (some) government software be open source. *TechCrunch, July*, 2016.

D. Cole. 'We kill people based on metadata'. *The New York Review of Books, May*, 2014.

P. Dabkowski and Y. Gal. Real time image saliency for black box classifiers. *CoRR*, abs/1705.07857, 2017.

F. Doshi-Velez and B. Kim. Towards a rigorous science of interpretable machine learning. *CoRR*, abs/1702.08608, 2017.

A. Einstein. The common language of science. *Advancement of Science*, 2(5), 1942. URL https://www.youtube.com/watch?v=e3B5BC4rhAU. The URL is for a 1941 recording by Einstein to the British Association for the Advancement of Science.

A. Etzioni. Is transparency the best disinfectant? *Journal of Political Philosophy*, 18(4):389–404, 2010.

K. Evtimova, A. Drozdov, D. Kiela, and K. Cho. Emergent language in a multi-modal, multi-step referential game. *CoRR*, abs/1705.10369, 2017.

R. Ghani. You say you want transparency and interpretability? 2016. Blog entry http://www.rayidghani.com/you-say-you-want-transparency-and-interpretability.

T. Gilovich, K. Savitsky, and V. Medvec. The illusion of transparency: biased assessments of others' ability to read one's emotional states. *Journal of personality and social psychology*, 75(2):332, 1998.

D. Griffin and L. Ross. Subjective construal, social inference, and human misunderstanding. *Advances in experimental social psychology*, 24:319–359, 1991.

R. Hammond and R. Axelrod. The evolution of ethnocentrism. *Journal of Conflict Resolution*, 50(6):926–936, 2006. For video, see http://www-personal.umich.edu/~axe/vtmovie.htm.

B. Horling and V. Lesser. A survey of multi-agent organizational paradigms. *The Knowledge Engineering Review*, 19(4):281–316, 2004.

F. Kelly. Network routing. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 337(1647):343–367, 1991.

H. Lakkaraju, S. Bach, and J. Leskovec. Interpretable decision sets: A joint framework for description and prediction. In *KDD*, pages 1675–1684. ACM, 2016.

E. Langer, A. Blank, and B. Chanowitz. The mindlessness of ostensibly thoughtful action: The role of "placebic" information in interpersonal interaction. *Journal of personality and social psychology*, 36(6):635–642, 1978.

N. Lawrence. Living together: Mind and machine intelligence. *CoRR*, abs/1705.07996, 2017.

J. Leibo, V. Zambaldi, M. Lanctot, J. Marecki, and T. Graepel. Multi-agent reinforcement learning in sequential social dilemmas. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, pages 464–473, 2017. Blog entry https://deepmind.com/blog/understanding-agent-cooperation/.

J. Lerner and P. Tetlock. Accounting for the effects of accountability. *Psychological bulletin*, 125(2):255, 1999.

E. Levine and M. Schweitzer. Prosocial lies: When deception breeds trust. *Organizational Behavior and Human Decision Processes*, 126:88–106, 2015.

Z. Lipton. The mythos of model interpretability. *CoRR*, abs/1606.03490, 2016.

R. Lowe, Y. Wu, A. Tamar, J. Harb, P. Abbeel, and I. Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. *CoRR*, abs/1706.02275, 2017. Blog entry https://blog.openai.com/learning-to-cooperate-compete-and-communicate/.

L. Lowenstein. Financial transparency and corporate governance: you manage what you measure. *Columbia Law Review*, 96(5): 1335–1362, 1996.

R. McAllister, Y. Gal, A. Kendall, M. van der Wilk, A. Shah, R. Cipolla, and A. Weller. Concrete problems for autonomous vehicle safety: Advantages of Bayesian deep learning. In *To appear in International Joint Conference on Artificial Intelligence (IJCAI)*, 2017.

I. Mordatch and P. Abbeel. Emergence of grounded compositional language in multi-agent populations. *CoRR*, abs/1703.04908, 2017.

R. Mortier, H. Haddadi, T. Henderson, D. McAuley, and J. Crowcroft. Human-data interaction: the human face of the data-driven society. *Social Science Research Network*, 2014.

A. Peled. When transparency and collaboration collide: The USA open data program. *Journal of the American society for information science and technology*, 62(11):2085–2094, 2011.

A. Prat. The wrong kind of transparency. *American Economic Review*, 95(3):862–877, 2005.

M. Ribeiro, S. Singh, and C. Guestrin. Why should I trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144. ACM, 2016.

T. Roughgarden. Selfish routing and the price of anarchy. Technical report, Stanford University Dept of Computer Science, 2006. URL http://www.dtic.mil/get-tr-doc/pdf?AD=ADA637949.

D. Solove. Privacy and power: Computer databases and metaphors for information privacy. *Stanford Law Review*, 53(6):1393–1462, 2001.

D. Solove. Why privacy matters even if you have 'nothing to hide'. *Chronicle of Higher Education*, 15, 2011.

D. Sridhar and R. Batniji. Misfinancing global health: a case for transparency in disbursements and decision making. *The Lancet*, 372(9644):1185–1191, 2008.

N. Tishby, F. Pereira, and W. Bialek. The information bottleneck method. In *Allerton Conference on Communication, Control, and Computing*, 1999.

T. Vishwanath and D. Kaufmann. Toward transparency: New approaches and their application to financial markets. *The World Bank Research Observer*, 16(1):41–57, 2001.

R. Wiley. The evolution of communication: information and manipulation. *Animal behaviour*, 2:156–189, 1983.

L. Zintgraf, T. Cohen, T. Adel, and M. Welling. Visualizing deep neural network decisions: Prediction difference analysis. In *ICLR*, 2017.