Beyond Transparency: A Proposed Framework for Accountability in Decision-Making AI Systems

Janelle Berscheid (University of Saskatchewan; j.berscheid@usask.ca)
Francois Roewer-Despres (University of Toronto; francoisrd@cs.toronto.edu)

DOI: 10.1145/3340470.3340476

Abstract

Transparency in decision-making AI systems can only become actionable in practice when all stakeholders share responsibility for validating outcomes. We propose a three-party regulatory framework that incentivizes collaborative development in the AI ecosystem and guarantees fairness and accountability are not merely afterthoughts in high-impact domains.

Introduction

Decision-making AI systems are becoming commonplace due to recent and rapid advances in computer hardware, machine learning algorithms, and an explosion in data availability. Increasingly, AI is being deployed to make life-altering decisions, such as those involving health, freedom, security, finance, and livelihood. However, the public is apprehensive about automated decision-making: the majority of Americans in a Pew Research Center survey were opposed to current applications of decision-making systems (Smith, 2018). Chief among their worries were concerns regarding the potential bias and unfairness of these systems, as well as skepticism regarding the validity of the decisions being made. Research on the social impact of AI, such as from New York University's AI Now Institute, corroborates these concerns, arguing that automated decision-making systems are often deployed untested and without accountability measures or processes for appeal (Whittaker et al., 2018).

Such technologies often create a gap between private and social cost. The motivating example for this essay was the COMPAS recidivism algorithm, developed by a company named Northpointe, which came under fire after a ProPublica investigation found that its results were racially biased and that jurisdictions were misapplying the results during trials, resulting in defendants incorrectly being labeled as potentially violent re-offenders (Angwin & Larson, 2016).

Stifling innovation is undesirable, but unchecked deployment of high-impact decision-making systems is damaging to the long-term health of the AI ecosystem. The public backlash will likely only intensify as the inevitable failures of unvetted systems come to light. Properly vetted systems have the potential to save time, money, and even lives, so finding remedies to the negative aspects of AI is critical to ensuring the public is a stakeholder in the AI ecosystem rather than an adversary.

Properly-designed AI systems present a unique opportunity to make transparent decisions by circumventing human fallibility. In fact, the framework we are proposing holds AI decision-makers to far more stringent standards than could ever be applied to human decision-makers. For example, psychological studies demonstrate human tendencies to make decisions subconsciously and then consciously rationalizing them post-hoc (Soon, Brass, Heinze, & Haynes, 2008), or to be prefer one decision over another based on framing of the situation (Tversky & Kahneman, 1981).

Ensuring accountability when mistakes are made is step towards addressing the concerns surrounding bias, fairness, and validity. Transparent decision-making—where the reasoning behind an AI's decision is made clear, interpretable, and auditable—is often proposed as a solution to the problem of biased or invalid AI systems. Indeed, transparency addresses many of these concerns; however, the gap between AI developers and other parties, such as the public and policymakers, cannot be closed by transparency alone. For example, the recent testimony of Facebook CEO Mark Zuckerberg before

Congress regarding the exploitation of Facebook users' data increased awareness surrounding this data misuse, but also revealed that many Congresspeople lack even a basic understanding of the technology (Kang, Kaplan, & Fandos, 2018). As a result, many pertinent questions were not asked, leading many to view it as a missed opportunity (Freedland, 2018).

Thus, we argue that transparency alone is not a sufficient requirement to produce accountability and reassure stakeholders: additional elements are needed to make transparency actionable in practice. For the public to become comfortable with autonomous decision-making systems, there needs to be a sense that a person can take remedial actions against an AI system if wronged. We identify two requirements to meeting such demands: disclosure of the existence of an AI in a decision-making process to the public, coupled with an appeal process for the AI's decisions, and a validated scope for the AI, including the use cases for which it been tested and its limitations.

These components suggest a need for a closer relationship between those who develop AI systems and those who create policies for their deployment. We propose a regulatory framework for AI systems that captures this need for communication into the development and deployment process itself. We envision establishing a pair of documents that ensure these disclosure and scope requirements are fulfilled prior. The developer's document publicly declares the scope of their system and how it has been validated, while the client document explicates the disclosure and appeal policies surrounding the deployment of a developer's system into a particular domain. In addition, a formally-established relationship between the developers and clients helps reduce cases where these two parties try to pin a system's failure on the other. In turn, this would facilitate the open sharing of knowledge and the cooperative development of fair procedures.

Definitions

For the purposes of this article, we want to make a careful distinction between the parties we have termed "developer", "client", and "subject".

Developer

An entity or person that has created an AI system for the purpose of real-world deployment. This may include, but is not limited to, those who designed the system, those who trained and tested the model at the core of the system, and those who developed the data pipeline.

Client

An entity or person that intends to deploy a developer's AI system as part of a decision-making process. This process may include other human or automated elements, as well as integrate AI systems from multiple developers. Typically, the client will be an institution (such as a charity, hospital, or government body) or a private company.

Subject

An entity or person that is the target of a client's deployed AI system. Subjects include patients, inmates, and consumers, or even private companies.

This distinction reflects a separation of concerns: developers are primarily concerned with the technical components of a system; clients are primarily concerned with the use of a system in relation to the subjects for a specific context. Further, this separation allows a developer to license their system to more than one client. It remains possible for the developer and client to be the same entity.

In addition, we want to clarify the types of AI systems being considered.

AI System

A system or process that incorporates AI in some form or another for the purpose of decision-making.

In this article, we will be considering requirements *only* for decision-making AI systems. We not do consider any AI systems that interact with, or perhaps mimic humans, as these may have a different set of requirements. We also focus our discussion on systems in high-impact areas, which may include, but are not limited to, health services, financial institutions, justice systems, aid programs, and civic initiatives. AI systems in lower-impact areas, such as recommender systems for consumer platforms, are less urgently in need of regulation, though these systems could benefit from the framework we describe as well. We also exempt AI research from our discussion, instead choosing to focus on systems intended for real-world deployment.

Transparency

We define transparency as making a decision that is in some way explainable, that has its inner workings available for review, and is not proprietary. This differs from disclosure, which is ‘transparency in the use of a model’ (rather than ‘transparency in the prediction of a model’). In other words, the use of a model can be disclosed, but the model itself might not be interpretable.

Implemented properly, transparency can help us examine erroneous assumptions made about the data, and even let human operators correct bad features in the model (Ribeiro, Singh, & Guestrin, 2016). Transparency also allows us to examine current and possible points of failure in a system, as well as monitor system performance to ensure the quality of decision-making does not degrade over time. In addition, transparency can help us verify that public stakeholders’ criteria, such as fairness or the removal of demographic bias, are being met.

Currently, many decision-making systems are neither transparent nor easily interpretable. Deep learning techniques, in particular, spur this concern, since neural networks are often characterized as a black box whose decision-making process is completely opaque. Explainable AI (XAI) research seeks to develop techniques to shine light on these systems, yet many fundamental problems remain open. Even the definitions of terms such as “transparency”, “interpretability”, and “explainability” are difficult to establish and even harder to unify across different input domains.

In addition, the mechanisms by which humans interpret these explanations, especially in relation to natural human subjectivity and bias, has yet to be understood. A recent survey of submissions from the International Joint Conferences on Artificial Intelligence workshop in XAI and found that knowledge from fields such as behavioural research and social science were rarely referenced, and the evaluations of any explanations that were produced did not generally include behavioural experiments (Miller, Howe, & Sonenberg, 2017). Behavioural analyses may be necessary to address subtler issues related to transparency. For example, the adequacy of a particular type of explanation may be dependent on the human interpreter. Indeed, doctors and patients use different explanatory models when interpreting a medical decision (Good, 1993).

Furthermore, hidden feedback loops, where multiple automated decision-making systems influence each other’s inputs over time, may also interfere with the system’s decision-making in a way that is difficult for a transparency algorithm to identify (Sculley et al., 2015). Transparency in this context may reveal a change over time, but cannot identify the influence of the other systems on the data collected, and thus on the model’s decision-making ability.

Despite these unsolved issues, transparency remains an integral component for accountability in AI. Exposing the inner workings of a model to external review helps foster trust in decision-making systems. However, transparency is not directly actionable in deployed systems unless an additional framework is in place to ensure subjects can use transparent explanations to hold developers and clients accountable when wronged.

Disclosure

For AI systems to be transparent, their use must be made known, rather than remaining a hidden part of a decision-making process. Applications that fail to disclose the use of AI systems create a power imbalance, where those unaware of its use are not in a position to question or challenge the figures making crucial decisions on their behalf. This leads to hidden biases and a lack of accountability, as well as creating confusion when problems do arise with these automated systems (Diallo, 2018).

Therefore, we argue that a person significantly impacted by a decision-making system should have the presence of this system clearly disclosed to them. This disclosure could be verbal or—in formal cases—form-based, requiring acknowledgement on the part of the subject. Many application domains of decision-making systems already require paperwork, such as health or legal systems, so viable disclosure channels already exist. A distinction can be made in disclosure between AI-*made* decisions, where the system makes a ruling without any human oversight, and AI-*assisted* decisions, where the system makes recommendations that are reviewed by a human as part of the final decision. Making this distinction could help diagnose points of failure in a decision-making process. First, whenever an AI-assisted process leads to biased or unfair decisions, measures such as retraining the human reviewers may be part of the solution. Second, an AI-assisted decision-making process in which the human reviewer always defers to the AI system's recommendation, may be treated as an effectively AI-made decision, which may violate the disclosure agreement.

In addition, disclosure will help developers identify stakeholders in the decision-making process who may otherwise be overlooked. For example, developers of a recidivism algorithm may think to consult prisoners in their requirements gathering, but may neglect to consult at-risk communities, which ought be consulted to create a fair and unbiased development process. Mandatory disclosure of AI use in their software would create an opportunity for these forgotten stakeholders to make their voices heard through an appeal process, and would encourage developers to expand their definition of impacted communities and stakeholders in future developments, creating a more community-conscious AI development process.

As alluded to, disclosure goes hand-in-hand with the additional requirement that autonomous decisions are appealable. An appeal process that is obscured or hidden cannot be described as fair, as it creates barriers limiting the participation of wronged subjects. When subjected to autonomous decisions, people ought to know not only that AI is employed, but also how its decisions can be examined or appealed. Specifying the details of the appeal process is outside the purview of this article, but we suggest that the process should be inclusive to all affected subjects, and have a minimal burden of entry so that members of marginalized communities could reasonably be expected to go through with the process.

Undoubtedly, this disclosure process would cause friction in the adoption of AI applications by the public. Affected individuals would likely challenge decisions, ask questions, and make appeals more frequently when the presence of AI in a decision-making process is publicized. This should be viewed as a positive long-term effect, rather than as a barrier to innovation and progress: overlooked stakeholders in these systems will be able to engage in a dialogue with the developers, clients, and institutions. Because the long-term health of AI-enabled technologies will ultimately depend on the public's trust and acceptance of these systems, AI developers should be responsible for winning the confidence of both regulating bodies and the general public regarding the efficacy, safety, and fairness of their systems. This will enable a more iterative and democratic process in AI development and deployment.

Validated Scope

Clearly defining the scope of autonomous decision-making systems protects subjects against spurious claims made by developers and clients, and protects developers against misapplication or misuse of their systems by clients. In addition, clearly-defined scope would help narrow the discovery phase, as well as reduce the decision-making burden, of legal cases, audits, and appeals of systems.

In life-altering application domains, care must be taken to ensure that automated systems do not introduce systematic biases and harmful side effects. Medical devices and drugs undergo heavy regulation to prevent unverified claims; why should life-altering autonomous decision-making systems not bear a similar burden? People should not be guinea pigs for algorithms deployed untested in real-world settings, no matter how promising the application, or how well the technology has worked elsewhere. While AI systems should not be limited to a single scope, nor prevented from being applied outside of their originally-intended context, a system's efficacy ought to be re-validated in each new deployment context.

Requiring a validated scope for autonomous decision-making systems would place the burden of proof of their system's efficacy on developers wishing to enter the market, and on clients wishing to transfer an existing system to a new market, rather than on wronged subjects trying to prove out-of-scope use. Given the power that decision-making systems can have over livelihoods, providing evidence that the system is functioning as intended seems obvious, yet stories continue to emerge of developers and clients rushing to deploy biased or largely-untested decision-making systems. ProPublica's investigation into Northpointe's COMPAS algorithm found that it was only about 20% accurate at identifying violent reoffenders; the figure for non-violent reoffenders was just 61%. Worse, the system was twice as likely to rule unfavourably for black offenders than white offenders (Angwin & Larson, 2016).

Additionally, some recent systems are deployed based on spurious, pseudoscientific claims, such as Predictim's rating potential babysitters for "disrespectful attitude" (Merchant, 2018), or companies claiming to determine personality traits and even "criminality" from facial features (Storm, 2016) (y Arcas, Mitchell, & Todorov, 2017). Requiring systems to validate the scope of their application before deployment causes such premises to fall apart. Attempting to define a metric by which to evaluate disrespectfulness may reveal inherent biases, which the AI will subsequently learn. Such applications, which are designed to prey on fear rather than provide a truly beneficial service, will be forced to either move towards evidence-based validation and metrics, or to explicitly state that their system is an elaborate placebo with no real predictive power, which will lessen their appeal.

Evaluating the use cases under which an AI system performs well will also bring to light its limitations. Though popular culture pushes the narrative that AI is rapidly approaching general intelligence, the current reality is that AI achieves high performance only at very narrow tasks, and does not yet generalize to other tasks. This leads to fragile, brittle systems, as demonstrated by the effectiveness of adversarial attacks (Szegedy et al., 2013).

Without a clearly-defined scope, clients without AI expertise may misunderstand how a developer's system is intended to be used, or may misinterpret the AI's decision. For example, the Northpointe COMPAS recidivism algorithm was originally designed to inform treatment, but, unbeknownst to the developer, was being used for sentencing: in Wisconsin, a judge overturned a plea deal after viewing a defendant's COMPAS risk score (Angwin & Larson, 2016). Whether due to fragility, unstated limitations, or the nature of the data, decision-making systems may work in one situation but not another, yet clients who fail to understand these issues may take systems which are fair, unbiased, and valid in one context, and unwittingly deploy them in another context which violates one of these conditions.

A clearly-defined scope can also help protect developers from charges of misuse of personal data. Regulation surrounding the use of these data are tightening, as seen with the European Union's (EU's) new General Data Protection Regulation (GDPR, 2016). A developer that collected data under a certain pretense to build a model with a specific purpose may be at least partially liable to violations of that pretense if a client uses the model for another purpose. However, if developers explicate the purpose of their system ahead of time, the liability for misapplication rests entirely with the clients.

Proposed Framework

To bring these requirements together in a way that places the onus on developers and clients to jointly create fair, unbiased, accountable systems, we propose a formal regulation system that requires a pair of documents to be filed for any autonomous decision-making system to be deployed in a high-impact application domain.

The first document, referred to as an AI Validation Document (AIVD), is developer-filed and concerns transparency and validated scope: it defines one or more contexts for which the system was developed, outlines the claims made with respect to the system in each context, demonstrate how these claims were validated, and explains how the system's decisions can be interpreted in the case of an audit. The second document, referred to as a

Deployment Disclosure Document (DDD), is client-filed and concerns disclosure and auditability: it identifies the process for deploying a developer's AI in validated contexts and the specific terms surrounding disclosure and appeal of the AI's decisions. If this deployment is legally challenged (e.g. through the appeal of a decision that is alleged to be biased or unfair), these documents can help determine fault, if any. Though they would be filed separately, requiring both documents incentivizes developers and clients to communicate carefully regarding the accountability surrounding any particular deployment, and to understand both the technical workings of the AI, as well as the domain-specific policies and challenges related to the deployment context.

Specifically, an AIVD consists of defining:

- the intended purpose of the system, and
- the conditions under which it can be safely used for the intended purpose
- how those use cases have been tested and what metrics have been used to validate them, including how the results have been checked for bias
- known limitations of those use cases
- a description of the data used to train the system, including when, where, and how it has been sourced
- measures undertaken to track the model's development, including all versions of the model that may be deployed, and the specifics of how the data was used to train and validate each version
- how results of the system may be interpreted

This avoids problems such as Northpointe's COMPAS having an unvetted bias against black offenders, as such bias would need to be identified in COMPAS's AIVD, which would result in an undeployable system.

This document would have to be filed first, to establish the validity of the application in at least one context. In addition, an AIVD could be amended over time to accommodate new use cases whenever the developer can provide sufficient evidence supporting an extension of scope.

Once an AIVD has been approved, clients wishing to use one or more developer applications as part of a larger process or system would need to file a DDD containing:

- the purpose of the decision-making process to be created
- the reference number of the AIVD(s) being deployed
- the specific context under which each referenced system will be deployed within the larger application
- the process for disclosing the use of AI to subjects of the decision-making process
- the process by which a subject can investigate or appeal a decision
- the process by which the system's decisions will be traced and linked to the individual models involved in the decision-making process, including the assignment of responsibility between interacting models, or models making recommendations to a human reviewer

This avoids problems such as Northpointe's COMPAS being used by courts in sentencing, which was never intended by Northpointe and thus would not have been in the COMPAS's AIVD. Then, any courts trying to misapply COMPAS would have their DDD application denied.

One AIVD could be associated with many different DDDs, across a wide variety of countries and contexts, provided the processes described in the DDD fall within the validated scope outlined in the AIVD. This incentivizes developers to create large, reliable, and robust AI systems, since the marginal cost of an additional deployment is small compared to the initial cost of developing the system. This incentive also suggests that an international treaty or body setting the standards for evaluating these systems and their deployment is favourable, as it affords economic benefits to filing in member nations, akin to the system of international agreements surrounding intellectual property.

However, international cooperation is not integral to the framework. Instead, the AIVD and DDD can be filed on a national basis, with a national (or regional, as in the EU) body of experts reviewing AIVD and DDD applications. Approving AIVDs requires significant expertise in various domains, including AI, and AIVDs will be subject to more scrutiny and longer evaluation times. In contrast, smaller groups of legal and domain experts are sufficient to evaluate DDDs in a timely fashion.

In the spirit of promoting transparency and accountability throughout, both documents should contain highly-technical, legal sections for expert review and simple, nontechnical sections for subject review. In other words, the documents themselves must be transparent. This helps prevent the common scenario in which users blindly agree to certain documents, such software end-user license agreements, because they are too complicated and lengthy to be read and understood.

Concerns surrounding the stifling of innovation are apparent with the introduction of these documents. As such, we propose that decision-making systems deployed in lower- or questionable-impact domains would not be required to file these documents, though developers and clients using AI in these areas would still be incentivized to file for the additional legal protection. This raises an issue regarding the definition of the impact level of a system. After all, seemingly benign systems, such as recommender systems on social media, can potentially have a huge impact on the general public's views and perception of reality (Mozur, 2018), even though this impact may be difficult to quantify, especially on the individual level. Still, we believe requiring AIVDs and DDDs only for high-impact domains achieves the best tradeoff between innovation and responsible development. In particular, it helps mitigate any potential stifling of innovation in the startup sphere—where a small team may not have the necessary data accessibility or legal expertise required to file AIVDs and DDDs—unless the startup is innovating in a clearly high-impact domain, in which case provisions for dealing with accountability ought to factor into their business model from the start. AI researchers would also be exempt from filing these documents, regardless of the impact level of their research, as the use of AI in this context is accounted for through the proper informed consent of research subjects, as well as the research ethics approval process.

Outside of a research setting, certain clients may need to engage in pilot projects before formally deploying a decision-making system. In this case, a precursor to a DDD is to be filed, which describes the attempted purpose, a timeline for the completion of the pilot phase of the project (after which a full DDD must be filed), and the process for disclosure. Disclosure in pilot projects must additionally include provisions for soliciting and incorporating feedback and concerns from subjects. This ensures that experimental or fringe ideas involve a significant amount of shared decision-making between clients and subjects.

Benefits of a Two-Phase Framework

We argue that this proposal encourages developers to share knowledge with clients, and to engage with the consequences of their systems, while also ensuring clients have a framework with which to evaluate and question developers regarding any planned deployment. AI expertise is scarce compared to demand (Perry, 2018), and many clients outside of the information technology domain—which may include institutions such as hospitals, justice systems, and civic organizations—may not be able to acquire or retain such expertise. Regulation in the form of a DDD will push clients to choose vetted developers, who have filed AIVDs, instead of those making baseless claims. Hype around the power of big data and machine learning may engender blind trust in non-expert clients; this document system may raise awareness regarding the efficacy of AI systems, as clients must evaluate the AIVDs of developers to file their own DDD. Both clients and policymakers would be encouraged to see AI systems as procedures that may have flaws and should therefore be examined and challenged. As AI failures become more pronounced in the public eye, developers pushing AI without a AIVD, even in a lower-impact domain, may be seen as a liability, encouraging more thoughtfully-developed and carefully-tested AI.

Under this document system, profit-driven developers have an incentive to help non-expert clients define their disclosure, tracing and appeal processes for each of their AIVDs, in order to quickly file DDDs for as many clients as possible. In addition, developers wanting to work with clients outside the scope of their AIVD would have to devise mechanisms for efficiently validating these new use cases, leading to much-needed innovation in this area, while simultaneously promoting accountability. This also incentivizes developers to concern themselves with the real-world implications of their systems on subjects. Essentially, both developers and clients are incentivized to proactively establish fair and accountable applications of AI systems, instead of considering it an afterthought until bad press comes to light.

Formally-defined accountability documents that are widely recognized also have the potential to raise public awareness around the technical, legal and ethical issues of decision-making AI. Laypeople, who will likely be the subjects of decision-making systems, need to be informed about how these systems may be behaving (or misbehaving) and about their rights to appeal automated decisions. Much in the same way that common knowledge of copyright and patents brings awareness to issues of intellectual property, AIVDs and DDDs being commonly recognized as a crucial step of AI deployment will help subjects be aware of the requirement that these systems disclose the use of AI, are transparent, and are validated. This, in turn, will allow subjects to openly question institutions deploying AI, to be more involved in the development of automated processes, and to hold these institutions accountable.

Conclusion

As decision-making AI systems are becoming ubiquitous, concerns surrounding bias, fairness, and accountability are mounting. Transparency in these systems is critical in reducing their potential harm. However, transparency alone is not actionable without additional requirements to close the gap between developers, clients, policymakers, and the public, since developing and deploying fair and accountable AI systems requires increased awareness of the strengths and limitations of automated decision-making. For instance, clients may be unaware of the limitations of these systems, or of what constitutes fair and ethical AI. In contrast, developers may not understand how to create effective policy surrounding their systems' use, or how to deploy them safely in novel and untested contexts.

Introducing disclosure and scope into the AI system's development and deployment process not only encourages proactive collaboration between both parties, but also ensures subjects are made aware, and can appeal the decisions, of the system. Unlike transparency, where may open problems remain, disclosure and scope are primarily policy-based, and can thus be feasibly implemented in all AI systems. These requirements serve not only to render transparency algorithms actionable once properly developed, but also to educate all stakeholders about the potential and the pitfalls surrounding AI systems.

We proposed a regulatory framework for AI systems that formalizes the disclosure and scope requirements in the form of two documents, AIVDs and DDDs. The double-document structure of this framework incentivizes a more collaborative AI ecosystem, as well as ensuring that fairness and validity are not afterthoughts in high-impact AI systems. Fostering this dialogue between all stakeholders in a decision-making process spurs innovation, and will help developers, clients, policymakers, and the public realize the potential that effective, fair, and accountable automated systems have.

References

Angwin, J., & Larson, J. (2016, May). *Machine Bias.* Retrieved 2019-01-09, from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Diallo, I. (2018, June). *The Machine Fired Me.* Retrieved 2019-01-02, from https://idiallo.com/blog/when-a-machine-fired-me

Freedland, J. (2018, April). Zuckerberg got off lightly. Why are politicians so bad at asking questions? *The Guardian*. Retrieved 2019-02-16, from https://www.theguardian.com/commentisfree/2018/apr/11/mark-zuckerberg-facebook-congress-senate

GDPR. (2016, May). Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec. *Official Journal of the European Union*, *L119*, 1–88.

Good, B. J. (1993). *Medicine, rationality and experience: an anthropological perspective*. Cambridge University Press.

Kang, C., Kaplan, T., & Fandos, N. (2018, April). Knowledge Gap Hinders Ability of Congress to Regulate Silicon Valley. *The New York Times*. Retrieved 2019-02-16, from https://www.nytimes.com/2018/04/12/business/congress-facebook-regulation.html

Merchant, B. (2018, December). *Predictim Claims Its AI Can Flag 'Risky' Babysitters. So I Tried It on the People Who Watch My Kids.* Retrieved 2018-12-27, from https://gizmodo.com/predictim-claims-its-ai-can-flag-risky-babysitters-so-1830913997

Miller, T., Howe, P., & Sonenberg, L. (2017, December). Explainable AI: Beware of Inmates Running the Asylum Or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences. *arXiv:1712.00547 [cs]*. Retrieved 2019-01-08, from http://arxiv.org/abs/1712.00547 (arXiv: 1712.00547)

Mozur, P. (2018, October). A Genocide Incited on Facebook, With Posts From Myanmar's Military. *The New York Times*. Retrieved 2019-01-11, from https://www.nytimes.com/2018/10/15/technology/myanmar-facebook-genocide.html

Perry, T. S. (2018, September). *Intel Execs Address the AI Talent Shortage, AI Education, and the "Cool" Factor.* Retrieved 2019-01-10, from https://spectrum.ieee.org/view-from-the-valley/robotics/artificial-intelligence/intel-execs-address-the-ai-talent-shortage-ai-education-and-the-cool-factor

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "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 - KDD '16* (pp. 1135–1144). San Francisco, California, USA: ACM Press. Retrieved 2019-01-08, from http://dl.acm.org/citation.cfm?doid=2939672.2939778 doi: 10.1145/2939672.2939778

Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., ... Dennison, D. (2015). Hidden technical debt in machine learning systems. In *Advances in neural information processing systems* (pp. 2503–2511).

Smith, A. (2018, November). *Public Attitudes Toward Computer Algorithms.* Retrieved 2019-01-05, from http://www.pewinternet.org/2018/11/16/attitudes-toward-algorithmic-decision-making/

Soon, C. S., Brass, M., Heinze, H.-J., & Haynes, J.-D. (2008, May). Unconscious determinants of free decisions in the human brain. *Nature Neuroscience*, *11*(5), 543–545. Retrieved 2019-02-16, from http://www.nature.com/articles/nn.2112 doi: 10.1038/nn.2112

Storm, D. (2016, May). *Faception can allegedly tell if you're a terrorist just by analyzing your face.* Retrieved 2019-01-01, from https://www.computerworld.com/article/3075339/security/faception-can-allegedly-tell-if-youre-a-terrorist-just-by-analyzing-your-face.html

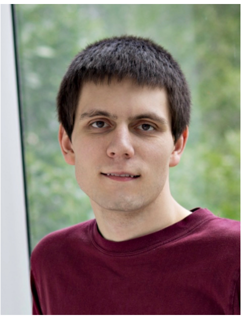
Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). Intriguing properties of neural networks. *arXiv:1312.6199 [cs]*. Retrieved 2019-01-08, from http://arxiv.org/abs/1312.6199 (arXiv: 1312.6199)

Tversky, A., & Kahneman, D. (1981, January). The framing of decisions and the psychology of choice. *Science*, *211*(4481), 453–458. Retrieved 2019-02-16, from http://www.sciencemag.org/cgi/doi/10.1126/science.7455683 doi: 10.1126/science.7455683

Whittaker, M., Crawford, K., Dobbe, R., Fried, G., Mathur, V., West, S. M., ... Schwartz, O. (2018, December). *AI Now Report 2018* (Tech. Rep.). AI Now Institute, New York University. Retrieved from https://ainowinstitute.org/AI_Now_2018_Report.pdf

y Arcas, B. A., Mitchell, M., & Todorov, A. (2017, May). *Physiognomy's New Clothes.* Retrieved 2019-01-07, from https://medium.com/@blaisea/physiognomys-new-clothes-f2d4b59fdd6a

Janelle Berscheid is a Master's student in the Computational Epidemiology and Public Health Informatics Lab (CEPHIL) at the University of Saskatchewan. Her current research focuses on health-related applications of data science and machine learning, and how these technologies can impact healthcare delivery.



Francois Roewer-Despres is a Master's student at the University of Toronto and a member of the Vector Institute for Artificial Intelligence. His research interests broadly cover Human-AI Interaction, including accountability and safety, autonomous social intelligence, dialogue systems, and reinforcement learning.