

1 **Draft NISTIR 8312**

2 **Four Principles of Explainable Artificial**  
3 **Intelligence**

4 P. Jonathon Phillips  
5 Carina A. Hahn  
6 Peter C. Fontana  
7 David A. Broniatowski  
8 Mark A. Przybocki

9 This draft publication is available free of charge from:  
10 <https://doi.org/10.6028/NIST.IR.8312-draft>

11 **NIST**  
**National Institute of**  
**Standards and Technology**  
U.S. Department of Commerce

12

## Draft NISTIR 8312

13

# Four Principles of Explainable Artificial Intelligence

14

15

P. Jonathon Phillips

16

Carina A. Hahn

17

Peter C. Fontana

18

*Information Access Division*

19

*Information Technology Laboratory*

20

David A. Broniatowski

21

*Information Technology Laboratory*

22

Mark A. Przybocki

23

*Information Access Division*

24

*Information Technology Laboratory*

25

This draft publication is available free of charge from:

26

<https://doi.org/10.6028/NIST.IR.8312-draft>

27

August 2020

28



29

U.S. Department of Commerce

30

*Wilbur L. Ross, Jr., Secretary*

31

National Institute of Standards and Technology

32

*Walter Copan, NIST Director and Undersecretary of Commerce for Standards and Technology*

33 **National Institute of Standards and Technology Interagency or Internal Report 8312**  
34 **24 pages (August 2020)**

35 This draft publication is available free of charge from:  
36 <https://doi.org/10.6028/NIST.IR.8312-draft>

37 Certain commercial entities, equipment, or materials may be identified in this document in  
38 order to describe an experimental procedure or concept adequately. Such identification is  
39 not intended to imply recommendation or endorsement by the National Institute of  
40 Standards and Technology, nor is it intended to imply that the entities, materials, or  
41 equipment are necessarily the best available for the purpose.

42 **Public comment period: August 17, 2020 through October 15, 2020**

43 National Institute of Standards and Technology  
44 100 Bureau Drive (Mail Stop 8940) Gaithersburg, Maryland 20899-2000  
45 Email: [explainable-AI@nist.gov](mailto:explainable-AI@nist.gov)

46 All comments will be made public and are subject to release under the Freedom of  
47 Information Act (FOIA).

48 Additional information on submitting comments can be found at  
49 <https://www.nist.gov/topics/artificial-intelligence/ai-foundational-research-explainability>.

50 **Trademark Information**

51 All trademarks and registered trademarks belong to their respective organizations.

52

## Call for Patent Claims

53 This public review includes a call for information on essential patent claims (claims  
54 whose use would be required for compliance with the guidance or requirements in this In-  
55 formation Technology Laboratory (ITL) draft publication). Such guidance and/or require-  
56 ments may be directly stated in this ITL Publication or by reference to another publication.  
57 This call also includes disclosure, where known, of the existence of pending U.S. or foreign  
58 patent applications relating to this ITL draft publication and of any relevant unexpired U.S.  
59 or foreign patents.

60 ITL may require from the patent holder, or a party authorized to make assurances on its  
61 behalf, in written or electronic form, either:

62 **a)** assurance in the form of a general disclaimer to the effect that such party does not hold  
63 and does not currently intend holding any essential patent claim(s); or

64 **b)** assurance that a license to such essential patent claim(s) will be made available to appli-  
65 cants desiring to utilize the license for the purpose of complying with the guidance  
66 or requirements in this ITL draft publication either:

67 **i.** under reasonable terms and conditions that are demonstrably free of any unfair  
68 discrimination; or

69 **ii.** without compensation and under reasonable terms and conditions that are demon-  
70 strably free of any unfair discrimination.

71 Such assurance shall indicate that the patent holder (or third party authorized to make assur-  
72 ances on its behalf) will include in any documents transferring ownership of patents subject  
73 to the assurance, provisions sufficient to ensure that the commitments in the assurance are  
74 binding on the transferee, and that the transferee will similarly include appropriate provi-  
75 sions in the event of future transfers with the goal of binding each successor-in-interest.

76 The assurance shall also indicate that it is intended to be binding on successors-in-  
77 interest regardless of whether such provisions are included in the relevant transfer docu-  
78 ments.

79 Such statements should be addressed to: [explainable-AI@nist.gov](mailto:explainable-AI@nist.gov)

80 **Abstract**

81 We introduce four principles for explainable artificial intelligence (AI) that comprise the  
82 fundamental properties for explainable AI systems. They were developed to encompass  
83 the multidisciplinary nature of explainable AI, including the fields of computer science,  
84 engineering, and psychology. Because one size fits all explanations do not exist, different  
85 users will require different types of explanations. We present five categories of explanation  
86 and summarize theories of explainable AI. We give an overview of the algorithms in the  
87 field that cover the major classes of explainable algorithms. As a baseline comparison, we  
88 assess how well explanations provided by people follow our four principles. This assess-  
89 ment provides insights to the challenges of designing explainable AI systems.

90

91 **Key words**

92 Artificial Intelligence (AI); explainable AI; trustworthy AI.

## Table of Contents

93		
94	<b>1 Introduction</b>	<b>1</b>
95	<b>2 Four Principles of Explainable AI</b>	<b>1</b>
96	2.1 Explanation	2
97	2.2 Meaningful	2
98	2.3 Explanation Accuracy	3
99	2.4 Knowledge Limits	4
100	<b>3 Types of Explanations</b>	<b>4</b>
101	<b>4 Overview of principles in the literature</b>	<b>6</b>
102	<b>5 Overview of Explainable AI Algorithms</b>	<b>7</b>
103	5.1 Self-Explainable Models	9
104	5.2 Global Explainable AI Algorithms	10
105	5.3 Per-Decision Explainable AI Algorithms	11
106	5.4 Adversarial Attacks on Explainability	12
107	<b>6 Humans as a Comparison Group for Explainable AI</b>	<b>12</b>
108	6.1 Explanation	13
109	6.2 Meaningful	13
110	6.3 Explanation Accuracy	14
111	6.4 Knowledge Limits	15
112	<b>7 Discussion and Conclusions</b>	<b>16</b>
113	<b>References</b>	<b>17</b>

## List of Figures

114		
115	Fig. 1 This figure shows length of response time versus explanation detail. We	
116	populate the figure with four illustrative cases: emergency weather alert,	
117	loan application, audit of a system, and debugging a system.	6

## 118 **1. Introduction**

119 With recent advances in artificial intelligence (AI), AI systems have become components of  
120 high-stakes decision processes. The nature of these decisions has spurred a drive to create  
121 algorithms, methods, and techniques to accompany outputs from AI systems with expla-  
122 nations. This drive is motivated in part by laws and regulations which state that decisions,  
123 including those from automated systems, provide information about the logic behind those  
124 decisions<sup>1</sup> and the desire to create trustworthy AI [30, 76, 89].

125 Based on these calls for explainable systems [40], it can be assumed that the failure to  
126 articulate the rationale for an answer can affect the level of trust users will grant that system.  
127 Suspicions that the system is biased or unfair can raise concerns about harm to oneself  
128 and to society [102]. This may slow societal acceptance and adoption of the technology,  
129 as members of the general public oftentimes place the burden of meeting societal goals  
130 on manufacturers and programmers themselves [27, 102]. Therefore, in terms of societal  
131 acceptance and trust, developers of AI systems may need to consider that multiple attributes  
132 of an AI system can influence public perception of the system.

133 Explainable AI is one of several properties that characterize trust in AI systems [83, 92].  
134 Other properties include resiliency, reliability, bias, and accountability. Usually, these terms  
135 are not defined in isolation, but as a part or set of principles or pillars. The definitions vary  
136 by author, and they focus on the norms that society expects AI systems to follow. For this  
137 paper, we state four principles encompassing the core concepts of explainable AI. These  
138 are informed by research from the fields of computer science, engineering, and psychology.  
139 In considering aspects across these fields, this report provides a set of contributions. First,  
140 we articulate the four principles of explainable AI. From a computer science perspective,  
141 we place existing explainable AI algorithms and systems into the context of these four prin-  
142 ciples. From a psychological perspective, we investigate how well people’s explanations  
143 follow our four principles. This provides a baseline comparison for progress in explainable  
144 AI.

145 Although these principles may affect the methods in which algorithms operate to meet  
146 explainable AI goals, the focus of the concepts is not algorithmic methods or computations  
147 themselves. Rather, we outline a set of principles that organize and review existing work in  
148 explainable AI and guide future research directions for the field. These principles support  
149 the foundation of policy considerations, safety, acceptance by society, and other aspects of  
150 AI technology.

## 151 **2. Four Principles of Explainable AI**

152 We present four fundamental principles for explainable AI systems. These principles are  
153 heavily influenced by considering the AI system’s interaction with the human recipient of  
154 the information. The requirements of the given situation, the task at hand, and the consumer

---

<sup>1</sup>The Fair Credit Reporting Act (FCRA) and the European Union (E.U.) General Data Protection Regulation (GDPR) Article 13.

155 will all influence the type of explanation deemed appropriate for the situation. These situa-  
156 tions can include, but are not limited to, regulator and legal requirements, quality control of  
157 an AI system, and customer relations. Our four principles are intended to capture a broad  
158 set of motivations, reasons, and perspectives.

159 Before proceeding with the principles, we need to define a key term, the *output* of an AI  
160 system. The output is the result of a query to an AI system. The output of a system varies by  
161 task. A loan application is an example where the output is a decision: approved or denied.  
162 For a recommendation system, the output could be a list of recommended movies. For a  
163 grammar checking system, the output is grammatical errors and recommended corrections.

164 Briefly, our four principles of explainable AI are:

165 **Explanation:** Systems deliver accompanying evidence or reason(s) for all outputs.

166 **Meaningful:** Systems provide explanations that are understandable to individual users.

167 **Explanation Accuracy:** The explanation correctly reflects the system’s process for gen-  
168 erating the output.

169 **Knowledge Limits:** The system only operates under conditions for which it was designed  
170 or when the system reaches a sufficient confidence in its output.

171 These are defined and contextualized in more detail below.

## 172 2.1 Explanation

173 The *Explanation* principle obligates AI systems to supply evidence, support, or reasoning  
174 for each output. By itself, this principle does not require that the evidence be correct, infor-  
175 mative, or intelligible; it merely states that a system is capable of providing an explanation.  
176 A body of ongoing work currently seeks to develop and validate explainable AI methods.  
177 An overview of these efforts is provided in Section 5. A variety of strategies and tools  
178 are currently being deployed and developed. This principle does not impose any metric  
179 of quality on those explanations. The Meaningful and Explanation Accuracy principles  
180 provide a framework for evaluating explanations.

## 181 2.2 Meaningful

182 A system fulfills the *Meaningful* principle if the recipient understands the system’s ex-  
183 planations. Generally, this principle is fulfilled if a user can understand the explanation,  
184 and/or it is useful to complete a task. This principle does not imply that the explanation is  
185 one size fits all. Multiple groups of users for a system may require different explanations.  
186 The Meaningful principle allows for explanations which are tailored to each of the user  
187 groups. Groups may be defined broadly as the developers of a system vs. end-users of a  
188 system; lawyers/judges vs. juries; etc. The goals and desiderata for these groups may vary.  
189 For example, what is meaningful to a forensic practitioner may be different than what is  
190 meaningful to a juror [31].



191 This principle also allows for tailored explanations at the level of the individual. Two  
192 humans viewing the same AI system’s output will not necessarily interpret it the same way  
193 for a variety of reasons. One reason is that a person’s prior knowledge and experiences in-  
194 fluence their decisions [45]. Another reason is that psychological differences among people  
195 may influence how they interpret an explanation and what type of explanations they find  
196 meaningful [10, 61]. Thus, different users may take different meanings from identical AI  
197 explanations. The tailoring of an explanation to user groups and individuals may not be  
198 static over time. As people gain experience with a task, what they consider a meaningful  
199 explanation will likely change [10, 35, 57, 72, 73]. Therefore, meaningfulness is influ-  
200 enced by a combination of the AI system’s explanation and a person’s prior knowledge,  
201 experiences, and mental processes.

202 All of the factors that influence meaningfulness contribute to the difficulty in model-  
203 ing the interface between AI and humans. Developing systems that produce meaningful  
204 explanations need to account for both computational and human factors [22, 58].

### 205 **2.3 Explanation Accuracy**

206 Together, the Explanation and Meaningful principles only call for a system to produce ex-  
207 planations that are meaningful to a user community. These two principles do not require  
208 that a system delivers an explanation that correctly reflects a system’s process for gen-  
209 erating its output. The *Explanation Accuracy* principle imposes accuracy on a system’s  
210 explanations.

211 Explanation accuracy is a distinct concept from decision accuracy. For decision tasks,  
212 decision accuracy refers to whether the system’s judgment is correct or incorrect. Re-  
213 gardless of the system’s decision accuracy, the corresponding explanation may or may not  
214 accurately describe *how* the system came to its conclusion. Researchers in AI have de-  
215 veloped standard measures of algorithm and system accuracy [13, 18, 33, 64–66, 71, 79].  
216 While there exist these established decision accuracy metrics, researchers are in the process  
217 of developing performance metrics for explanation accuracy [2, 16, 97].

218 Similarly to the Meaningful principle, this principle allows for different explanation  
219 accuracy metrics for different groups and individuals. Some users will require simple ex-  
220 planations that succinctly focus on the critical point(s) but lack nuances that are necessary  
221 to completely characterize the algorithm’s process for generating its output. However,  
222 these nuances may only be meaningful to experts. This highlights the point that explana-  
223 tion accuracy and meaningfulness need not overlap. A detailed explanation may be highly  
224 accurate but sacrifice how meaningful it is to certain audiences. Overall, a system may  
225 be considered more explainable if it can generate more than one type of of explanation.  
226 Because of these different levels of explanation, the metrics used to evaluate the accuracy  
227 of an explanation may not be universal or absolute.

## 228 2.4 Knowledge Limits

229 The previous principles implicitly assume that a system is operating within its knowledge  
230 limits. This *Knowledge Limits* principle states that systems identify cases they were not  
231 designed or approved to operate, or their answers are not reliable. By identifying and  
232 declaring knowledge limits, this practice safeguards answers so that a judgment is not pro-  
233 vided when it may be inappropriate to do so. The Knowledge Limits Principle can increase  
234 trust in a system by preventing misleading, dangerous, or unjust decisions or outputs.

235 There are two ways a system can reach its knowledge limits. First, the question can be  
236 outside the domain of the system. For example, in a system built to classify bird species, a  
237 user may input an image of an apple. The system could return an answer to indicate that it  
238 could not find any birds in the input image; therefore, the system cannot provide an answer.  
239 This is both an answer and an explanation. In the second way a knowledge limit can be  
240 reached, the confidence of the most likely answer may be too low, depending on an internal  
241 confidence threshold. For example, for a bird classification system, the input image of a  
242 bird may be too blurry to determine its species. In this case, the system may recognize that  
243 the image is of a bird, but that the image is of low quality. An example output may be: “I  
244 found a bird in the image, but the image quality is too low to identify it.”

## 245 3. Types of Explanations

246 Explanations will vary depending on their consumer. Some explanations will be simple,  
247 while others will be detailed and could require training or expertise to fully understand. To  
248 illustrate the range of explanation, we describe five categories of explanations that build on  
249 the work in the literature [6, 26, 98]. The categories described below were not designed to  
250 be exhaustive.

251 **User benefit:** This type of explanation is designed to inform a user about an output. For  
252 example, the explanation could provide the reason a loan application was approved  
253 or denied to the applicant.

254 **Societal acceptance:** This type of explanation is designed to generate trust and acceptance  
255 by society. For example, if an unexpected output is provided by the system, the  
256 explanation may help users understand why this output was generated. It may also  
257 provide an increased sense of comfort in the system if the rationale can be provided  
258 (e.g., [1]).

259 **Regulatory and compliance:** This type of explanation assists with audits for compliance  
260 with regulations, safety standards, etc. The audience of the explanation may include  
261 a user who requires significant detail (e.g., a safety regulator) and the user interacting  
262 with the system (e.g., a developer). Examples may include the developer or auditor  
263 of a self-driving car. This may also include explanations to evaluate the output of a  
264 forensic examination after an airplane crash.

265 **System development:** This type of explanation assists or facilitates developing, improv-  
266 ing, debugging, and maintaining of an AI algorithm or system. Consumers of this  
267 category includes technical staff, product managers, and executives. This category  
268 includes the users requiring significant detail and users interacting with the system.  
269 For example, this may include the technical staff debugging a vision algorithm with  
270 a Gradient-Weighted Class Activation Mapping (GRAD-CAM) based tool [82].

271 **Owner benefit:** This type of explanation benefits the operator of a system. An example  
272 is a recommendation system that lists movies or videos to watch and explains the  
273 selection based on previous viewed items. A system recommends a movie and ex-  
274 plains this choice by stating “here is a movie to watch because you liked these other  
275 movies.” If the user trusts the explanation, the owner benefits because that person  
276 continues watching movies on their service.

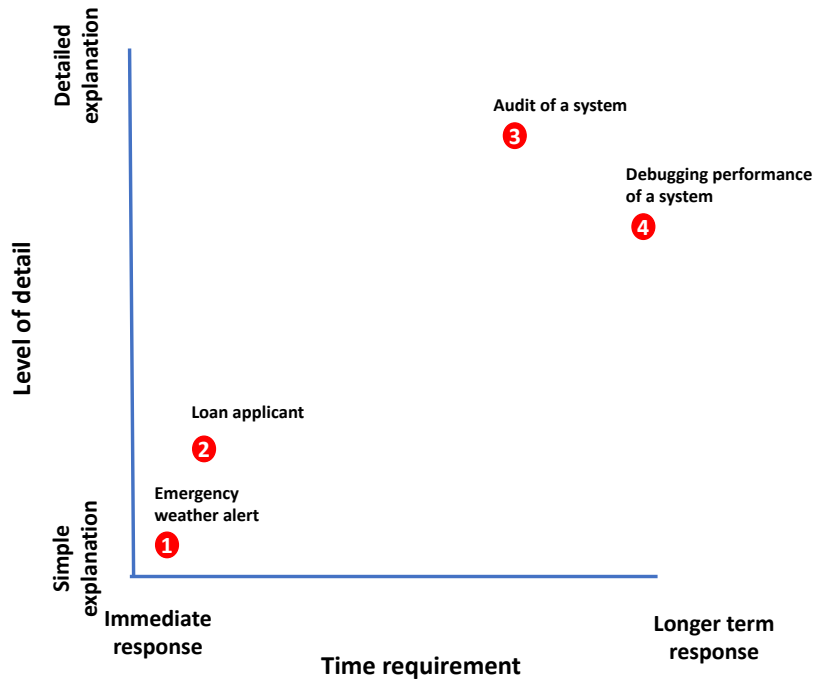
277 Categories of this nature are also discussed in more detail in Bhatt et al. [6], Hall et al.  
278 [26], Weller [98]. Bhatt et al. [6] mentions in their use cases that the explanations are  
279 usually used by the algorithm developers to debug the models. Bhatt et al. [6] interviews  
280 30 individuals on how their organizations use explainable AI. They use explainable AI in  
281 a variety of applications, including object detection and sentiment analysis. Hall et al.  
282 [26] proposes best practices on how to use explainable AI algorithms. They summarize  
283 their recommendations into implementation guidelines: design explanations to enable un-  
284 derstanding, learn how explainable AI can be exploited for nefarious purposes, augment  
285 surrogate models with direct explanations, and for high-stakes decisions, provided expla-  
286 nations must be highly interpretable. In Caruana et al. [11], the authors developed an  
287 explainable AI model and used it to both determine and explain pneumonia risk in a patient  
288 data set and 30-day readmission risk in another patient data set.

289 From a practical perspective, explanations can be characterized by the amount of time  
290 the consumer of the explanation has to respond to the information and the level of detail in  
291 an explanation. Figure 1 captures the relationship between time requirements and explana-  
292 tion detail. The horizontal axis represents the *time requirement* a user has to respond to a  
293 situation. The time requirement axis addresses situations ranging from those that require  
294 immediate responses to those that permit a longer evaluation. The vertical axis represents  
295 the *level of detail* in the explanation. This axis addresses situations related to the level of  
296 detail the consumer or user will require. At one end of the explanation, an explanation  
297 is not required or a simple explanation will be sufficient. For example, in response to an  
298 emergency weather alert, the consumer must act immediately, and the explanation needs to  
299 be simple and straightforward. A current weather alert from the National Weather Service,  
300 “Tornado Warning: Take Action!”<sup>2</sup>, operates as both an alert and a simple explanation. The  
301 alert is to “Take Action” with the simple explanation of “Tornado Warning.” Explanations  
302 for debugging could fall at the other end of the time requirement and level of detail spec-  
303 trum. The explanation could include information on the internal steps of a system, and it

---

<sup>2</sup><https://www.weather.gov/safety/tornado-ww>

304 could take the audience time to examine the explanation and decide on their next actions.  
 305 Two additional examples were placed on Figure 1: loan applications and audit of a sys-  
 306 tem. The response to a loan application is generally quick and the explanation provides  
 307 greater detail than a weather alert. The response time and explanation detail for an audit of  
 308 a system could be similar to debugging a system.



**Fig. 1.** This figure shows length of response time versus explanation detail. We populate the figure with four illustrative cases: emergency weather alert, loan application, audit of a system, and debugging a system.

309 Explanations will need to fulfill a variety of requirements and needs, which will depend  
 310 on the tasks and users. The five categories of explanations illustrate the range and types of  
 311 explanations and points to the need for flexibility in addressing the scope of systems that  
 312 require explanations.

#### 313 4. Overview of principles in the literature

314 Theories and properties of explainable AI have been discussed from different perspectives,  
 315 with commonalities and differences across these points of view [16, 22, 53, 77, 78, 98].

316 Lipton [53] divides explainable techniques into two broad categories: transparent and  
 317 post-hoc interpretability. Lipton [53] defines a transparent explanation as reflecting to some  
 318 degree how a system came to its output. A subclass is simulatability, which requires that

319 a person can grasp the entire model. This implies that explanations will reflect the inner  
320 workings of a system. Their post-hoc explanations “often do not elucidate precisely how  
321 a model works, they may nonetheless confer useful information for practitioners and end  
322 users of machine learning.” For example, the bird is a cardinal because it is similar to  
323 cardinals in the training set.

324 Rudin [77] and Rudin and Radin [78] argue that models for high-stakes decision must  
325 provide explanations that reveal their inner workings. They claim that deep neural networks  
326 are inherently black-boxes and should be avoided for high-stakes decisions.

327 Wachter et al. [97] argue that explanations do need to meet the explanation accuracy  
328 property. They claim that counterfactual explanations are sufficient. “A counterfactual ex-  
329 planation of a prediction describes the smallest change to the feature values that changes the  
330 prediction to a predefined output [59];” e.g., if you had arrived to the platform 15 minutes  
331 earlier, you would have caught the train. Counterfactual explanations do not necessarily  
332 reveal the inner workings of a system. This property allows counterfactual explanations to  
333 protect intellectual property.

334 Gilpin et al. [22] defines a set of concepts for explainable AI and provides an outline  
335 of current approaches. In their survey, Gilpin et al. [22] take a similar stance to Rudin [77]  
336 and Rudin and Radin [78] in their set of “foundational concepts” for explainability. Similar  
337 to the meaningful and explanation accuracy principles in our current work, Gilpin et al.  
338 [22] propose that explanations should allow for a trade-off between their interpretability  
339 and completeness. However, they state that trade-offs must not obscure key limitations of  
340 a system.

341 Doshi-Velez and Kim [16] address the critical question: measuring if explanations are  
342 meaningful for users or consumers. They present a framework for a science to measure the  
343 efficiency of explanations. This paper discusses factors that are required to begin testing  
344 interpretability of explainable systems. This highlights the gap between these principles as  
345 a concept and creating metrics and evaluation methods.

346 Across these viewpoints, there exist both commonalities and disagreement. Similar to  
347 our four principles, commonalities include concepts which distinguish between the exist-  
348 ence of an explanation, how meaningful it is, and how accurate or complete it is. Although  
349 disagreements remain, these perspectives provide guidance for development of explainable  
350 systems. A key disagreement between philosophies is the relative importance of explana-  
351 tion meaningfulness and accuracy. These disagreements highlight the difficulty in balanc-  
352 ing multiple principles simultaneously. Context of the application, community and user  
353 requirements, and the specific task will drive the importance of each principle.

## 354 **5. Overview of Explainable AI Algorithms**

355 Researchers have developed different algorithms to explain AI systems. Sometimes, the  
356 algorithms themselves provide the explanation (satisfying Principle 1). The most common  
357 of these explanations are *self-explainable models*, where the models themselves are the  
358 provided explanation. These models are self-explaining algorithms, where viewing and

359 querying the models provide an explanation. We describe these algorithms in Section 5.1.  
360 There are algorithms that provide explanations for themselves without directly providing  
361 the model details. One such example is Class Activation Mappings (CAM) [105], which are  
362 system-specific explanations that can explain some convolutional neural networks. How-  
363 ever, researchers generalized these algorithms so that they can not only explain the original  
364 system but also explain other systems. These generalized algorithms form the next two  
365 types of explanations: global explainable AI algorithms and per-decision explainable AI  
366 algorithms. For instance, GRAD-CAM is a generalization of CAM that can provide the  
367 explanation of CAM but to any convolutional neural network [82].

368 A *global explanation* produces a model that approximates the non-interpretable model.  
369 We describe these algorithms in Section 5.2. *Per-decision explanations* provide a separate  
370 explanation for each decision. Per-decision explanations are considered *local explanations*.  
371 We describe per-decision explanations in Section 5.3. A particular type of per-decision ex-  
372 planation is a *counterfactual*[97], which is an explanation saying “if the input were this  
373 new input instead, the system would have made a different decision.” In these explana-  
374 tions, although there are often many widely-differing instances that all are counterfactuals,  
375 a counterfactual explanation usually provides a single instance. This means that even if  
376 there are many different possible ways that the instance could be changed to result in the  
377 system providing the decision, only one of those instances is provided as the explanation.  
378 The hope is the instance is as similar as possible to the input with the exception that the  
379 system makes a different decision. Because counterfactual explanations are per-decision  
380 explanations, they are also described in Section 5.3.

381 Self-explainable models of machine learning systems themselves can be used as global  
382 explanations (since the models explain themselves). Likewise, many global explanations  
383 (including self-explainable models) can also be used to generate per-decision explanations.  
384 The coefficient weights of the features of an input in a regression model and the flow of a  
385 decision through a decision tree both serve as per-decision explanations. Models that do not  
386 provide an explanation or provide an explanation that a user does not consider meaningful  
387 enough will sometimes seek an explanation from an alternate algorithm, thus encouraging  
388 the development of global and per-decision explanations. Furthermore, global explanations  
389 are harder to generate than per-decision explanations because per-decision explanations  
390 only require an understanding of a single decision.

391 With these explainable algorithms, developers wish for the explanations to be mean-  
392 ingful to users (Principle 2). In the computer science literature this is often labelled as  
393 *interpretable*. Often, developers self-proclaim their algorithm explanations to be mean-  
394 ingful. However, others will use measurements such as human simulatability [46], which  
395 measure whether a human can correctly take an input and with the model, correctly identify  
396 the model’s prediction.

397 Although the explanation accuracy is important (Principle 3), it is often only measured  
398 for self-explainable models. For these types of models, the model’s decision accuracy  
399 (see Section 2.3) is the measure of the explanation accuracy. However, there is limited  
400 research measuring explanation accuracy. Adebayo et al. [2] evaluate explanation accu-

401 racy of saliency pixel explanations for deep neural networks by measuring the amount the  
402 explanation changes relative to how the trained models differ.

403 To our knowledge there is limited work on developing algorithms that understand their  
404 knowledge limits (Principle 4) and declare when a validly-formatted data input is out of  
405 the system’s scope. However, algorithms often give real-valued outputs rather than hard  
406 decisions, which reflect the algorithms’ confidences in their predictions.

## 407 5.1 Self-Explainable Models

408 Machine Learning Algorithms include Decision Trees and Linear and Logistic Regression.  
409 Although these simple models are explanations themselves, they are often not always ac-  
410 curate, especially if used without much pre-processing. Consequently, there has been work  
411 in developing more accurate models that themselves are explanations. Authors developing  
412 models will often label these models as interpretable, which we refer to as meaningful.  
413 Rudin [77] argues that using meaningful models that explain themselves are the best way  
414 to produce explainable models, arguing that separately-produced explanations of black-box  
415 models (or even single decisions of black-box models) may not be faithful to what the orig-  
416 inal model computes. This claim is that explanations often have low explanation accuracy  
417 if those explanations are not the models themselves. Although many sources discuss an  
418 accuracy-interpretability trade-off, Rudin and Radin [78] disagrees, with the belief that no  
419 such trade-off exists for high-stakes decisions.

420 One line of research works on producing improvements on the standard decision trees,  
421 sometimes represented as a nested sequence of “if-then-else” rules, called *decision lists*  
422 [47]. In addition to being inaccurate, Lakkaraju et al. [47] claims that the nesting makes the  
423 rules hard to interpret, and develops *Decision Sets*, which are a sequence of “if-then” rules  
424 with one default “else” at the end, where each clause is a conjunction of conditions. How-  
425 ever, Lakkaraju and Rudin [50] produces decision lists with improved accuracy. Lakkaraju  
426 et al. [47] measure the interpretability of the decision sets by measuring metrics on the  
427 model: the number of rules, the number of the largest rules, the overlap of the rules (how  
428 many instances are classified in more than one if-then rule). The last “else” guarantees that  
429 every instance is classified. [49] explores decision lists with at most one customized nesting  
430 to further improve accuracy while still being meaningful according to their measures. Bert-  
431 simas and Dunn [5] produce a variant of decision trees, called *optimal classification trees*,  
432 that split on mixed integer constraints involving multiple variables. These trees focus on  
433 preserving the meaningfulness of decision trees but greatly improving their classification  
434 accuracy. [55] produce another variant of a more accurate decision tree, called an addi-  
435 tive tree, that combines elements of decision tress and gradient boosting to produce more  
436 accurate trees. A Bayesian variant of decision lists that was studied for meaningfulness  
437 is Bayesian Rule Lists [51], where they add a Bayesian credible interval estimate to each  
438 decision rule. Bayesian credible intervals are the Bayesian analog to confidence intervals.  
439 Kuhn et al. [42] produces a model that tries to find combinations of features that either  
440 exclude a class or specifically identify a particular class. Each set of combinations could

441 be viewed as a clause of a decision set rule.

442 Models, including linear models such as linear and logistic regression are considered  
443 to be explanations of system decisions. One interpretation is using the weights of the coef-  
444 ficients to indicate the importance of features. They are sometimes considered inaccurate  
445 when the data is not believed to be linear. One measure of the ease of understanding of a re-  
446 gression model is the number of non-zero coefficients. One way to encourage a regression  
447 model to limit the number of features is to *regularize* it with the *lasso*, which penalizes the  
448 model for using more features [32], incorporating a trade-off for accuracy and meaningful-  
449 ness in the training objective function. Although this and other regularization strategies are  
450 also used to prevent overfitting in many models including deep neural networks, regulariza-  
451 tion is one technique to make models sparser, and thus believed to be more understandable.  
452 Poursabzi-Sangdeh et al. [69] considers regression models meaningful and aims to measure  
453 the value the model coefficients provide to human users trying to use the model. Caruana  
454 et al. [11] also treats the more general class of these models, Generalized Additive Mod-  
455 els with Pairwise Interactions (GA2M), as understandable models and applies them to a  
456 healthcare case study.

457 Another self-explainable algorithm involves learning *prototypes*, or representative sam-  
458 ples of each class, to better understand the algorithm. Models learn and produce prototypes.  
459 With these prototypes, the model outputs the class as a weighted combination of the proto-  
460 types. Although these prototypes do work on tabular data, Kim et al. [38], Li et al. [52] use  
461 this approach for classification on image data sets.

## 462 5.2 Global Explainable AI Algorithms

463 Global Explainable AI Algorithms are an approach that treat the AI algorithm as a black-  
464 box that can be queried and produce a model that explains the algorithm. Depending on  
465 what the global model is, it can then be used to produce per-decision explanations.

466 One such global explainable AI Algorithms is SHAP (SHapley Additive exPlanations)  
467 [56]. SHAP provides a global per-feature importance for a regression problem by convert-  
468 ing it to a coalitional game from game theory. In coalitional games, there are  $n$  players that  
469 can team up in different ways to form coalitions and share a payoff depending on which  
470 players team up (often the total payoff is largest when all players team up). After play-  
471 ers receive a payoff, they must divide the payoff between themselves. One way to divide  
472 payoffs with desirable mathematical properties is to give each player their Shapley value  
473 as their individual payoff. SHAP treats the regression outputs of a system as a coalitional  
474 game where the target is the payoff and each feature is a player that either participates in or  
475 does not participate in the coalition with the other features for each row. SHAP then com-  
476 putes the Shapley values for each feature, and uses those values as the feature importance  
477 values. See [20] for more information on Shapley values and coalitional games.

478 In deep neural networks, one such global algorithm is TCAV (Testing with Concept Ac-  
479 tivation Vectors) [107]. TCAV wishes to explain a neural network in a more user-friendly  
480 way by representing the neural network state as a linear weighting of human-friendly con-



481 cepts, called Concept Activation Vectors (CAVs). TCAV was applied to explain image  
482 classification algorithms through learning CAVs including color, to see how colors influ-  
483 enced the image classifier’s decisions.

484 Two visualizations used to provide global explanations are Partial Dependence Plots  
485 (PDPs) and Individual Conditional Expectation (ICE) [60, 104]. The partial dependence  
486 plot shows the marginal change of the predicted response when the feature (value of that  
487 specific data column or component) changes. PDPs are useful for determining if a relation-  
488 ship between a feature and the response is linear or more complex [60]. The ICE curves are  
489 finer-grained and show the marginal effect of the change in one feature for each instance  
490 of the data. ICE curves are useful to check if the relationship visualized in the PCP is the  
491 same across all ICE curves, and can help identify potential interactions.

### 492 5.3 Per-Decision Explainable AI Algorithms

493 Per-decision explainable AI algorithms take both a black-box model that can be queried and  
494 a single decision of that model, and explain why the model made that particular decision.  
495 These explanations differ from global explanations in that the explanation is not required  
496 to generalize to other decisions.

497 One such algorithm is LIME (Local Interpretable Model-Agnostic Explainer) [74].  
498 LIME takes a decision, and by querying nearby points, builds an interpretable model that  
499 represents the local decision, and then uses that model to provide per-feature explanations.  
500 The default model chosen is logistic regression. For images, LIME breaks each image into  
501 superpixels, and then queries the model with a random search space where it varies which  
502 superpixels are omitted and replaced with all black (or a color of the user’s choice).

503 Another popular type of local explanations are counterfactuals. A *counterfactual* expla-  
504 nation is an alternate system input where the system’s decision on that input differs from the  
505 provided input. Good counterfactuals answer the question “what is the minimum amount  
506 an input would need to change for the system to change its decision on that input?” Wachter  
507 et al. [97] measures how good counterfactual explanations are by measuring how far away  
508 the counterfactual is from the original data point, measuring this distance as the Manhattan  
509 distance of features after normalizing each feature by its median absolute deviation. Ustun  
510 et al. [96] develop a counterfactual explanation of logistic (or linear) regression models.  
511 Counterfactuals are represented as the amounts of specific features to change. They further  
512 refine their counterfactual explanations by distinguishing which features can be changed,  
513 which ones cannot, and which ones can only be changed under certain conditions.

514 An additional local explanation in Koh and Liang [39] takes a decision and produces  
515 an estimate of the influence of each training data point on that particular decision.

516 Another popular type of local explanations for problems on image data are *saliency*  
517 *pixels*. Saliency pixels color each pixel depending on how much that pixel contributes to  
518 the classification decision. One of the first saliency algorithms is Class Activation Maps  
519 (CAM) [105]. A popular saliency pixel algorithm that enhanced CAM is GRAD-CAM  
520 [82]. GRAD-CAM generalized CAM so that it can explain any convolutional network.

521 A variety of saliency pixel explanation algorithms are compared on for their explanation  
522 accuracy in Adebayo et al. [2].

#### 523 **5.4 Adversarial Attacks on Explainability**

524 Explanation accuracy (Principle 3) is an important component of explanations. Sometimes,  
525 if an explanation does not have 100 percent explanation accuracy, it can be exploited by  
526 adversaries who manipulate a classifier’s output on small perturbations of an input to hide  
527 the biases of a system. First, Lakkaraju and Bastani [48] observes that even if an expla-  
528 nation can mimic the predictions of the black box, that this is insufficient for explanation  
529 accuracy and such systems can produce explanations that may mislead users. An approach  
530 to generate misleading explanations is demonstrated in Slack et al. [84]. They do this by  
531 producing a scaffolding around a given classifier that matches the classification on all in-  
532 put data instances but changes outputs for small perturbations of input points, which can  
533 obfuscate global system behavior when only queried locally. This means that if the sys-  
534 tem is anticipating being explained by a tool such as LIME that gives similar instances to  
535 training set instances as inputs, the system will develop an alternative protocol to decide  
536 those instances that differs from how they will classify trials in the training and test sets.  
537 This can mislead the explainer by anticipating which trials the system might be asked to  
538 classify. Another similar approach is demonstrated in Aivodji et al. [3]. They fairwash a  
539 model by taking a black box model and produce an ensemble of interpretable models that  
540 approximate the original model but are much fairer, which then hide the unfairness of the  
541 original model. Another example of slightly perturbing a model to manipulate explanations  
542 is demonstrated in Dimanov et al. [14]. The ability for developers to cover up unfairness in  
543 black-box models is one of the several vulnerabilities of explainable AI discussed in Hall  
544 et al. [26].

### 545 **6. Humans as a Comparison Group for Explainable AI**

546 Up to this point, we have outlined core concepts of explainable AI and related work in  
547 the field of computer science. However, an explainable AI system consists of both an AI  
548 system and a human recipient. To effectively understand both components, and to provide  
549 a benchmark for explainable AI systems, we next overview the explainability of human-  
550 produced judgments and decisions. Independent of AI, humans operating alone also make  
551 high stakes decisions with expectation that they be explainable. For example, physicians,  
552 judges, lawyers, and forensic scientists make decisions that can affect large populations.  
553 In these cases, a human makes the decision and provides their conclusion along with the  
554 evidence supporting that conclusion as an explanation. How do these proffered explana-  
555 tions adhere to our four principles? We focused strictly on human explanations of their  
556 own judgments and decisions (e.g., “why did you arrive at this conclusion or choice?”), not  
557 of external events (e.g., “why is the sky blue?” or “why did an event occur?”). External  
558 events accompanied by explanations can be helpful for human reasoning and formulating

559 predictions [54]. This is consistent with a desire for explainable AI. However, as outlined  
560 in what follows, human-produced explanations for their own judgments, decisions, and  
561 conclusions are largely unreliable. Humans as a comparison group for explainable AI can  
562 inform the development of benchmark metrics for explainable AI systems; and lead to a  
563 better understanding of the dynamics of human-machine collaboration.

## 564 **6.1 Explanation**

565 This principle requires only that the system provides an explanation. In this section, we will  
566 focus on whether humans produce explanations of their own judgments and decisions and  
567 whether doing so is beneficial for the decision makers themselves. In Section 6.2, we will  
568 discuss whether human explanations are meaningful, and in Section 6.3, we will discuss  
569 the accuracy of those explanations.

570 Humans are able to produce a variety of explanation types [37, 53, 58]. However,  
571 producing verbal explanations can interfere with decision and reasoning processes [80, 81,  
572 100]. It is thought that as one gains expertise, the underlying processes become more  
573 automatic, outside of conscious awareness, and therefore, more difficult to explain verbally  
574 [17, 19, 44, 80]. This produces a similar tension which exists for AI itself: the desire for  
575 high accuracy are often thought to come with reductions in explainability (however, c.f.,  
576 [53]).

577 More generally, processes which occur with limited conscious awareness can be harmed  
578 by requiring the decision itself to be expressed explicitly. An example of this comes from  
579 lie detection. Lie detection based on explicitly judging whether or not a person is telling  
580 the truth or a lie is typically inaccurate [9, 88]. However, when judgments are provided  
581 via implicit categorization tasks, therefore bypassing an explicit judgment, lie detection  
582 accuracy can be improved [87, 88]. This suggests that lie detection may be a nonconscious  
583 process which is interrupted when forced to be made a conscious one.

584 Together these findings suggest that some assessments from humans may be more ac-  
585 curate when left automatic and implicit, compared to requiring an explicit judgment or  
586 explanation. Human judgments and decision making can oftentimes operate as a black-box  
587 [53], and interfering with this black-box process can be deleterious to the accuracy of a  
588 decision.

## 589 **6.2 Meaningful**

590 To meet this principle, the system provides explanations that are intelligible and under-  
591 standable. For this, we focused on the ability of humans to interpret how another human  
592 arrived at a conclusion. This concept can be defined operationally as: 1) whether the au-  
593 dience reaches the same conclusion as intended by the person providing the explanation  
594 and 2) whether the audience agrees with each other on what the conclusion is, based on an  
595 explanation.

596 One analogous case to explainable AI for human-to-human interaction is that of a foren-  
597 sic scientist explaining forensic evidence to laypeople (e.g., members of a jury). Currently,

598 there is a gap between the ways forensic scientists report results and the understanding of  
599 those results by laypeople (see Edmond et al. [17], Jackson et al. [31] for reviews). Jack-  
600 son et al. [31] extensively studied the types of evidence presented to juries and the ability  
601 for juries to understand that evidence. They found that most types of explanations from  
602 forensic scientists are misleading or prone to confusion. Therefore, they do not meet our  
603 internal criteria for being “meaningful.” A challenge for the field is learning how to improve  
604 explanations, and the proposed solutions do not always have consistent outcomes [31].

605 Complications for producing meaningful explanations for others include people expect-  
606 ing different explanation types, depending on the question at hand [58], context driving the  
607 formation of opinions [31], and individual differences in what is considered to be a satis-  
608 factory explanation [61]. Therefore, what is considered meaningful varies by context and  
609 across people.

### 610 **6.3 Explanation Accuracy**

611 This principle states that a system provides explanations which are faithful to the system’s  
612 process for generating the output. For humans, this is analogous to an explanation of one’s  
613 decision processes truly reflecting the mental processes behind that decision. In this sec-  
614 tion, we focused on this aspect only. An evaluation of the quality or coherence of the  
615 explanation falls outside of the scope of this principle.

616 For the type of introspection related to explanation accuracy, it is well-documented that  
617 although people often report their reasoning for decisions, this does not reliably reflect  
618 accurate or meaningful introspection [62, 70, 99]. This has been coined the “introspection  
619 illusion”: a term to indicate that information gained by looking inward to one’s mental  
620 contents is based on mistaken notions that doing so has value [70]. People fabricate reasons  
621 for their decisions, even those thought to be immutable, such as personally held opinions  
622 [24, 34, 99]. In fact, people’s conscious reasoning that is able to be verbalized does not  
623 seem to always occur before the expressed decision. Instead, evidence suggests that people  
624 make their decision and then apply reasons for those decisions *after* the fact [95]. From a  
625 neuroscience perspective, neural markers of a decision can occur up to 10 seconds before  
626 a person’s conscious awareness [85]. This finding suggests that decision making processes  
627 begin long before our conscious awareness.

628 People are largely unaware of their inability to introspect accurately. This is docu-  
629 mented through studies of “choice blindness” in which people do not accurately recall their  
630 prior decisions. Despite this inaccurate recollection, participants will provide reasons for  
631 making selections they never, in fact, made [24, 25, 34]. For studies that do not involve  
632 long-term memory, participants have also been shown to be unaware of the ways they eval-  
633 uate perceptual judgments. For example, people are inaccurate when reporting which facial  
634 features they use to determine someone’s identity [75, 93].

635 Based on our definition of explanation accuracy, these findings do not support the idea  
636 that humans reliably meet this criteria. As is the case with algorithms, human decision  
637 accuracy and explanation accuracy are distinct. For numerous tasks, humans can be highly

638 accurate but cannot verbalize their decision process.

#### 639 **6.4 Knowledge Limits**

640 This principle states that the system only operates under the conditions it was designed  
641 or that a provided output may not be reliable. For this principle, we narrowed down the  
642 broad field of *metacognition*, or thinking about one’s own thinking. Here, we focused on  
643 whether humans correctly assess their own ability and accuracy, and whether they know  
644 when to report that they do not know an answer. There are several ways to test whether  
645 people can evaluate their own knowledge limits. One method is to ask participants to  
646 predict how well they believe they performed or will perform on a task, relative to others  
647 (e.g., in what percentile will their scores fall relative to other task-takers). Another way to  
648 test the awareness of knowledge limits is to obtain a measure of their response confidence,  
649 with higher confidence indicating that a person believes with greater likelihood that they  
650 are correct.

651 As demonstrated by the well known Dunning-Kruger Effect [41], most people inaccurately  
652 estimate their own ability relative to others. A similar finding is that people, including  
653 experts, generally do not *predict* their own accuracy and ability well when asked  
654 to explicitly estimate performance [7, 8, 12, 28, 63]. However, a recent replication of the  
655 Dunning-Kruger Effect for face perception showed that, although people did not reliably  
656 predict their accuracy, their ability estimates varied accordingly with the task difficulty  
657 [106]. This suggests that although the exact value (e.g., predicted performance percentile  
658 relative to others, or predicted percent correct) may be erroneous, people can modulate the  
659 direction of their predicted performance appropriately (e.g., knowing a task was more or  
660 less difficult for them).

661 For a variety of judgments and decisions, people often know when they have made  
662 errors, even in the absence of feedback [103]. To use eyewitness testimony as a relevant  
663 example: although confidence and accuracy have repeatedly shown to be weakly related  
664 [86], a person’s confidence does predict their accuracy in the absence of “contamination”  
665 through the interrogation process and extended time between the event and the time of  
666 recollection [101]. Therefore, human shortcomings in assessing their knowledge limits are  
667 similar to those of producing explanations themselves. When asked explicitly to produce  
668 an explanation, these explanations can interfere with more automatic processes gained by  
669 expertise; they often do not accurately reflect the true cognitive processes. Likewise, as  
670 outlined in this section, when people are asked to explicitly predict or estimate their ability  
671 level relative to others, they are often inaccurate. However, when asked to assess their  
672 confidence for a given decision vs. this explicit judgment, people can gauge their accuracy  
673 at levels above chance. This suggests people do have insight into their own knowledge  
674 limits, although this insight can be limited or weak in some cases.

## 675 7. Discussion and Conclusions

676 We introduced four principles to encapsulate the fundamental elements for explainable AI  
677 systems. The principles provide a framework with which to address different components  
678 of an explainable system. These four principles are that the system produce an explanation,  
679 that the explanation be meaningful to humans, that the explanation reflects the system’s  
680 processes accurately, and that the system expresses its knowledge limits. There are differ-  
681 ent approaches and philosophies for developing and evaluating explainable AI. Computer  
682 science approaches tackle the problem of explainable AI from a variety of computational  
683 and graphical techniques and perspectives, which may lead to promising breakthroughs. A  
684 blossoming field puts humans at the forefront when considering the effectiveness of AI ex-  
685 planations and the human factors behind their effectiveness. Our four principles provide a  
686 multidisciplinary framework with which to explore this type of human-machine interaction.

687 The practical needs of the system will influence how these principles are addressed (or  
688 dismissed). With these needs in mind, the community will ultimately adapt and apply the  
689 four principles to capture a wide scope of applications. One example of adapting to meet  
690 practical requirements is illustrated by the trade-off between explanation detail and time  
691 constraints. These constraints highlight that certain scenarios require a brief, meaningful  
692 explanation to take priority over an accurate, detailed explanation. For example, emergency  
693 weather alerts need to be meaningful to the public but can lack an accurate explanation  
694 of how the system arrived at its conclusion. Other scenarios may require more detailed  
695 explanations but restrict meaningfulness to a specific user group; e.g., when auditing a  
696 model.

697 The focus of explainable AI has been to advance the capability of the systems to pro-  
698 duce a quality explanation. Here, we addressed whether humans can meet the same set of  
699 principles we set forth for AI. We showed that humans demonstrate only limited ability to  
700 meet the principles outlined here. This provides a benchmark with which to compare AI  
701 systems. In reflection of societal expectations, recent regulations have imposed a degree  
702 of accountability on AI systems via the requirement for explainable AI [1]. As advances  
703 are made in explainable AI, we may find that certain parts of AI systems are better able  
704 to meet societal expectations and goals compared to humans. By understanding the ex-  
705 plainability of both the AI system and the human in the human-machine collaboration, this  
706 opens the door to pursue implementations which incorporate the strengths of each, poten-  
707 tially improving explainability beyond the capability of either the human or AI system in  
708 isolation.

709 In this paper, we focused on a limited set of human factors related to explainable de-  
710 cisions. Much is to be learned and studied regarding the interaction between humans and  
711 explainable machines. Although beyond the scope of the current paper, in considering the  
712 interface between AI and humans, understanding general principles that drive human rea-  
713 soning and decision making may prove to be highly informative for the field of explainable  
714 AI [23]. For humans, there are general tendencies for preferring simpler and more general  
715 explanations [58]. However, as described earlier, there are individual differences in which

716 explanations are considered high quality. The context of the decision and the type of de-  
717 cision being made can influence this as well. Humans do not make decisions in isolation  
718 of other factors [45]. Without conscious awareness, people incorporate irrelevant infor-  
719 mation into a variety of decisions such as first impressions, personality trait judgments,  
720 and jury decisions [21, 29, 90, 91]. Even when provided identical information, the con-  
721 text, a person’s biases, and the way in which information is presented influences decisions  
722 [4, 15, 17, 23, 36, 43, 68, 94]. Considering these human factors within the context of  
723 explainable AI has only just begun.

724 To succeed in explainable AI, the community needs to study the interface between hu-  
725 mans and AI systems. Human-machine collaborations have shown to be highly effective  
726 in terms of accuracy [67]. There may be similar breakthroughs for AI explainability in  
727 human-machine collaborations. The principles defined here provide guidance and a phi-  
728 losophy for driving explainable AI toward a safer world by giving users a deeper under-  
729 standing into a system’s output. Meaningful and accurate explanations empower users to  
730 apply this information to adapt their behavior and/or appeal decisions. For developers and  
731 auditors, explanations equips them with the ability to improve, maintain, and deploy sys-  
732 tems as appropriate. Explainable AI contributes to the safe operation and trust of multiple  
733 facets of complex AI systems. The common framework and definitions under the four prin-  
734 ciples facilitate the evolution of explainable AI methods necessary for complex, real-world  
735 systems.

## 736 **Acknowledgments**

737 The authors thank Kristen Greene, Reva Schwartz, Brian Stanton, Amy Yates, and Jesse  
738 Zhang for their insightful comments and discussions.

## 739 **References**

- 740 [1] (2018). General Data Protection Regulation (GDPR).  
741 [2] Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., and Kim, B. (2018).  
742 Sanity Checks for Saliency Maps. In Bengio, S., Wallach, H., Larochelle, H., Grau-  
743 man, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information*  
744 *Processing Systems 31*, pages 9505–9515. Curran Associates, Inc.  
745 [3] Aivodji, U., Arai, H., Fortineau, O., Gambs, S., Hara, S., and Tapp, A. (2019). Fair-  
746 washing: the risk of rationalization. In *International Conference on Machine Learning*,  
747 pages 161–170. ISSN: 1938-7228 Section: Machine Learning.  
748 [4] Bertrand, M. and Mullainathan, S. (2004). Are Emily and Greg More Employable Than  
749 Lakisha and Jamal?: A Field Experiment on Labor Market Discrimination. *American*  
750 *Economic Review*, 94(4):991–1013.  
751 [5] Bertsimas, D. and Dunn, J. (2017). Optimal classification trees. *Machine Learning*,  
752 106(7):1039–1082.

- 753 [6] Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., Ghosh, J., Puri, R.,  
754 Moura, J. M., and Eckersley, P. (2020). Explainable machine learning in deployment.  
755 In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*,  
756 pages 648–657.
- 757 [7] Bindemann, M., Attard, J., and Johnston, R. A. (2014). Perceived ability and actual  
758 recognition accuracy for unfamiliar and famous faces. *Cogent Psychology*, 1(1).
- 759 [8] Bobak, A. K., Mileva, V. R., and Hancock, P. J. (2018). Facing the facts: Naive partici-  
760 pants have only moderate insight into their face recognition and face perception abilities.  
761 *Quarterly Journal of Experimental Psychology*, page 174702181877614.
- 762 [9] Bond, C. F. and DePaulo, B. M. (2006). Accuracy of Deception Judgments Character-  
763 izations of Deception. *Personality and Social Psychology Review*, 10(3):214–234.
- 764 [10] Broniatowski, D. A. and Reyna, V. F. (2018). A formal model of fuzzy-trace theory:  
765 Variations on framing effects and the Allais paradox. *Decision (Wash D C)*, 5(4):205–  
766 252.
- 767 [11] Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., and Elhadad, N. (2015). Intel-  
768 ligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-Day Read-  
769 mission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowl-  
770 edge Discovery and Data Mining, KDD '15*, pages 1721–1730, New York, NY, USA.  
771 Association for Computing Machinery. event-place: Sydney, NSW, Australia.
- 772 [12] Chi, M. (2006). Two approaches to the study of experts’ characteristics. In Ericsson,  
773 K., Charness, N., Feltovich, P., and Hoffman, R., editors, *The Cambridge Handbook  
774 of Expertise and Expert Performance*, chapter 2, pages 21–30. Cambridge University  
775 Press, Cambridge.
- 776 [13] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). ImageNet:  
777 A Large-Scale Hierarchical Image Database. In *IEEE Conference on Computer Vision  
778 and Pattern Recognition (CVPR)*.
- 779 [14] Dimanov, B., Bhatt, U., Jamnik, M., and Weller, A. (2020). You shouldn’t trust  
780 me: Learning models which conceal unfairness from multiple explanation methods. In  
781 *European Conference on Artificial Intelligence*.
- 782 [15] Doleac, J. L. and Stein, L. C. (2013). The visible hand: Race and online market  
783 outcomes. *The Economic Journal*, 123(572):F469–F492.
- 784 [16] Doshi-Velez, F. and Kim, B. (2017). Towards a rigorous science of interpretable  
785 machine learning. *arXiv preprint arXiv:1702.08608*.
- 786 [17] Edmond, G., Towler, A., Grows, B., Ribeiro, G., Found, B., White, D., Ballantyne,  
787 K., Searston, R. A., Thompson, M. B., Tangen, J. M., Kemp, R. I., and Martire, K.  
788 (2017). Thinking forensics: Cognitive science for forensic practitioners. *Science and  
789 Justice*, 57(2):144–154.
- 790 [18] Everingham, M., Eslami, S. M. A., Van Gool, L., Williams, C. K. I., Winn, J., and  
791 Zisserman, A. (2015). The pascal visual object classes challenge: A retrospective. *In-  
792 ternational Journal of Computer Vision*, 111(1):98–136.
- 793 [19] Fallshore, M. and Schooler, J. W. (1995). Verbal Vulnerability of Perceptual Ex-  
794 pertise. *Journal of Experimental Psychology: Learning, Memory, and Cognition*,



- 795 21(6):1608–1623.
- 796 [20] Ferguson, T. (2014). *Game Theory*. Second edition.
- 797 [21] Flowe, H. D. and Humphries, J. E. (2011). An examination of criminal face bias in a  
798 random sample of police lineups. *Applied Cognitive Psychology*, 25(2):265–273.
- 799 [22] Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., and Kagal, L. (2018). Ex-  
800 plaining explanations: An overview of interpretability of machine learning. *Proceedings*  
801 *- 2018 IEEE 5th International Conference on Data Science and Advanced Analytics,*  
802 *DSAA 2018*, pages 80–89.
- 803 [23] Google LLC (2019). AI Explanations Whitepaper. pages 1–28.
- 804 [24] Hall, L., Johansson, P., and Strandberg, T. (2012). Lifting the Veil of Morality: Choice  
805 Blindness and Attitude Reversals on a Self-Transforming Survey. *PLoS ONE*, 7(9).
- 806 [25] Hall, L., Johansson, P., Tärning, B., Sikström, S., and Deutgen, T. (2010). Magic at  
807 the marketplace: Choice blindness for the taste of jam and the smell of tea. *Cognition*,  
808 117(1):54–61.
- 809 [26] Hall, P., Gill, N., and Schmidt, N. (2019). Proposed guidelines for the responsible use  
810 of explainable machine learning.
- 811 [27] Haney, J. and Furman, S. (2019). Perceptions of Smart Home Privacy and Security  
812 Responsibility, Concerns, and Mitigations. *15th Symposium on Usable Privacy and*  
813 *Security*.
- 814 [28] Harvey, N. (1997). Confidence in judgment. *Trends in Cognitive Sciences*, 1(2):78–  
815 82.
- 816 [29] Hu, Y., Parde, C. J., Hill, M. Q., Mahmood, N., and O’Toole, A. J. (2018). First Im-  
817 pressions of Personality Traits From Body Shapes. *Psychological Science*, 29(12):1969–  
818 1983.
- 819 [30] IBM Research (Accessed July 8, 2020). Trusting AI. Available at [https://www.  
820 research.ibm.com/artificial-intelligence/trusted-ai/](https://www.research.ibm.com/artificial-intelligence/trusted-ai/).
- 821 [31] Jackson, G., Kaye, D. H., Neumann, C., Ranadive, A., and Reyna, V. F. (2015). Com-  
822 municating the Results of Forensic Science Examinations. Technical report.
- 823 [32] James, G., Witten, D., Hastie, T., and Tibshirani, R. (2014). *An Introduction to Sta-  
824 tistical Learning: with Applications in R*. Springer, New York, 1st edition in 2013,  
825 corrected 4th printing 2014 edition edition.
- 826 [33] Japkowicz, N. and Shah, M. (2014). *Evaluating Learning Algorithms A Classification  
827 Perspective*. Cambridge University Press.
- 828 [34] Johansson, P., Hall, L., Sikström, S., and Olsson, A. (2005). Failure to de-  
829 tect mismatches between intention and outcome in a simple decision task. *Science*,  
830 310(5745):116–119.
- 831 [35] Kahneman, D. (2011). *Thinking, Fast and Slow*. Farrar, Straus and Giroux, New  
832 York.
- 833 [36] Kassin, S. M., Dror, I. E., and Kukucka, J. (2013). The forensic confirmation bias:  
834 Problems, perspectives, and proposed solutions. *Journal of Applied Research in Memory  
835 and Cognition*, 2(1):42–52.
- 836 [37] Keil, F. C. (2006). Explanation and understanding. *Annual Review of Psychology*,

- 837 57:227–254.
- 838 [38] Kim, B., Rudin, C., and Shah, J. A. (2014). The Bayesian Case Model: A Generative  
839 Approach for Case-Based Reasoning and Prototype Classification. In Ghahramani, Z.,  
840 Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q., editors, *Advances in  
841 Neural Information Processing Systems 27*, pages 1952–1960. Curran Associates, Inc.
- 842 [39] Koh, P. W. and Liang, P. (2017). Understanding Black-Box Predictions via Influence  
843 Functions. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML’17, pages 1885–1894. JMLR.org. event-place: Sydney, NSW,  
844 Australia.  
845
- 846 [40] Kroll, J. A., Huey, J., Barocas, S., Felton, E. W., Reidenberg, J. R., Robinson, D. G.,  
847 and Yu, H. (2017). Accountable Algorithms. *University of Pennsylvania Law Review*,  
848 pages 633–705.
- 849 [41] Kruger, J. and Dunning, D. (1999). Unskilled and unaware of it: How difficulties  
850 in recognizing one’s own incompetence lead to inflated self-assessments. *Journal of  
851 Personality and Social Psychology*, 77(6):1121–1134.
- 852 [42] Kuhn, D. R., Kacker, R., Lei, Y., and Simos, D. E. (2020). Combinatorial Methods  
853 for Explainable AI. In *IWCT 2020*. Library Catalog: conf.researchr.org.
- 854 [43] Kukucka, J., Kassin, S. M., Zapf, P. A., and Dror, I. E. (2017). Cognitive Bias and  
855 Blindness: A Global Survey of Forensic Science Examiners. *Journal of Applied Re-  
856 search in Memory and Cognition*, 6(4):452–459.
- 857 [44] Kulatunga-Moruzy, C., Brooks, L. R., and Norman, G. R. (2004). Using compre-  
858 hensive feature lists to bias medical diagnosis. *Journal of Experimental Psychology:  
859 Learning Memory and Cognition*, 30(3):563–572.
- 860 [45] Kveraga, K., Ghuman, A. S., and Bar, M. (2007). Top-down prediction in the cogni-  
861 tive brain. *Brain and cognition*, 65(2):145–168.
- 862 [46] Lage, I., Chen, E., He, J., Narayanan, M., Kim, B., Gershman, S. J., and Doshi-Velez,  
863 F. (2019). Human Evaluation of Models Built for Interpretability. *Proceedings of the  
864 AAAI Conference on Human Computation and Crowdsourcing*, 7(1):59–67.
- 865 [47] Lakkaraju, H., Bach, S. H., and Leskovec, J. (2016). Interpretable Decision Sets:  
866 A Joint Framework for Description and Prediction. In *Proceedings of the 22nd ACM  
867 SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD  
868 ’16, pages 1675–1684, New York, NY, USA. Association for Computing Machinery.  
869 event-place: San Francisco, California, USA.
- 870 [48] Lakkaraju, H. and Bastani, O. (2020). “how do i fool you?”: Manipulating user trust  
871 via misleading black box explanations. In *Proceedings of the AAAI/ACM Conference on  
872 AI, Ethics, and Society*, AIES ’20, page 79–85, New York, NY, USA. Association for  
873 Computing Machinery.
- 874 [49] Lakkaraju, H., Kamar, E., Caruana, R., and Leskovec, J. (2019). Faithful and Cus-  
875 tomizable Explanations of Black Box Models. In *Proceedings of the 2019 AAAI/ACM  
876 Conference on AI, Ethics, and Society*, AIES ’19, pages 131–138, New York, NY, USA.  
877 Association for Computing Machinery. event-place: Honolulu, HI, USA.
- 878 [50] Lakkaraju, H. and Rudin, C. (2017). Learning Cost-Effective and Interpretable Treat-

- 879 ment Regimes. In *Artificial Intelligence and Statistics*, pages 166–175.
- 880 [51] Letham, B., Rudin, C., McCormick, T. H., and Madigan, D. (2015). Interpretable  
881 classifiers using rules and Bayesian analysis: Building a better stroke prediction model.  
882 *The Annals of Applied Statistics*, 9(3):1350–1371.
- 883 [52] Li, O., Liu, H., Chen, C., and Rudin, C. (2018). Deep Learning for Case-Based  
884 Reasoning Through Prototypes: A Neural Network That Explains Its Predictions. In  
885 *Thirty-Second AAAI Conference on Artificial Intelligence*.
- 886 [53] Lipton, Z. C. (2018). The mythos of model interpretability. *Communications of the*  
887 *ACM*, 61(10):36–43.
- 888 [54] Lombrozo, T. (2006). The structure and function of explanations. *Trends in Cognitive*  
889 *Sciences*, 10(10):464–470.
- 890 [55] Luna, J. M., Gennatas, E. D., Ungar, L. H., Eaton, E., Diffenderfer, E. S., Jensen, S. T.,  
891 Simone, C. B., Friedman, J. H., Solberg, T. D., and Valdes, G. (2019). Building more  
892 accurate decision trees with the additive tree. *Proceedings of the National Academy of*  
893 *Sciences*, 116(40):19887–19893.
- 894 [56] Lundberg, S. M. and Lee, S.-I. (2017). A Unified Approach to Interpreting Model  
895 Predictions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vish-  
896 wanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Sys-*  
897 *tems 30*, pages 4765–4774. Curran Associates, Inc.
- 898 [57] Marti, D. and Broniatowski, D. A. (2020). Does gist drive NASA experts’ design  
899 decisions? *Systems Engineering*, (May 2019):1–20.
- 900 [58] Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sci-  
901 ences. *Artificial Intelligence*, 267:1–38.
- 902 [59] Molnar, C. (2018). *Interpretable Machine Learning*.
- 903 [60] Molnar, C. (2019). *Interpretable Machine Learning*. @ChristophMolnar, online edi-  
904 tion edition.
- 905 [61] Mueller, S. T., Hoffman, R. R., Clancey, W., Emrey, A., and Klein, G. (2019). Ex-  
906 planation in Human-AI Systems: A Literature Meta-Review, Synopsis of Key Ideas  
907 and Publications, and Bibliography for Explainable AI. *arXiv:1902.01876 [cs]*. arXiv:  
908 1902.01876.
- 909 [62] Nisbett, R. E., Wilson, T. D., Kruger, M., Ross, L., Indeed, A., Bellows, N.,  
910 Cartwright, D., Goldman, A., Gurwitz, S., Lemley, R., London, H., and Markus, H.  
911 (1977). Telling more than we can know: Verbal reports on mental processes. *Psycho-*  
912 *logical Review*, 84(3).
- 913 [63] Oskamp, S. (1965). Overconfidence in case-study judgments. *Journal of Consulting*  
914 *Psychology*, 29(3):261–265.
- 915 [64] Phillips, P., Bowyer, K. W., Flynn, P. J., Liu, X., and Scruggs, W. T. (2008). The Iris  
916 Challenge Evaluation 2005. In *Second IEEE International Conference on Biometrics:*  
917 *Theory, Applications, and Systems*.
- 918 [65] Phillips, P. J., Moon, H., Rizvi, S., and Rauss, P. (2000). The FERET evaluation  
919 methodology for face-recognition algorithms. *IEEE Trans. PAMI*, 22:1090–1104.
- 920 [66] Phillips, P. J., Scruggs, W. T., O’Toole, A. J., Flynn, P. J., Bowyer, K. W., Schott,

- 921 C. L., and Sharpe, M. (2010). FRVT 2006 and ICE 2006 large-scale results. *IEEE*  
922 *Trans. PAMI*, 32(5):831–846.
- 923 [67] Phillips, P. J., Yates, A. N., Hu, Y., Hahn, C. A., Noyes, E., Jackson, K., Cavazos,  
924 J. G., Jeckeln, G., Ranjan, R., Sankaranarayanan, S., et al. (2018). Face recognition  
925 accuracy of forensic examiners, superrecognizers, and face recognition algorithms. *Pro-*  
926 *ceedings of the National Academy of Sciences*, 115(24):6171–6176.
- 927 [68] Pohl, R. F., editor (2004). *Cognitive illusions: A handbook on fallacies and biases in*  
928 *thinking, judgement and memory*. Psychology Press.
- 929 [69] Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Vaughan, J. W., and Wallach,  
930 H. (2019). Manipulating and Measuring Model Interpretability. *arXiv:1802.07810 [cs]*.  
931 arXiv: 1802.07810.
- 932 [70] Pronin, E. (2009). The introspection illusion. In *Advances in experimental social*  
933 *psychology*, pages 1–67. Elsevier.
- 934 [71] Przyboccki, M. A., Martin, A. F., and Le, A. N. (2007). Nist speaker recognition eval-  
935 uations utilizing the mixer corpora—2004, 2005, 2006. *IEEE Transactions on Audio,*  
936 *Speech, and Language Processing*, 15(7):1951–1959.
- 937 [72] Reyna, V. F. (2012). A new intuitionism: Meaning, memory, and development in  
938 Fuzzy-Trace Theory Valerie. *Judgment and Decision Making*, 7(3):332–359.
- 939 [73] Reyna, V. F. (2018). When Irrational Biases Are Smart: A Fuzzy-Trace Theory of  
940 Complex Decision Making. *Journal of Intelligence*, 6(2):29.
- 941 [74] Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "Why Should I Trust you?" Ex-  
942 plaining the Predictions of Any Classifier. In *KDD 2016: Proceedings of the 22nd ACM*  
943 *SIGKDD Conference on Knowledge Discovery and Data Mining*, San Francisco, CA,  
944 USA. ACM.
- 945 [75] Rice, A., Phillips, P. J., and O’Toole, A. J. (2013). The role of the face and body in  
946 unfamiliar person identification. *Applied Cognitive Psychology*, 27:761–768.
- 947 [76] Roach, J. (Accessed July 29, 2020). Microsoft responsible machine learning capabil-  
948 ities build trust in AI systems, developers say. Available at [https://blogs.microsoft.com/](https://blogs.microsoft.com/ai/azure-responsible-machine-learning/)  
949 [ai/azure-responsible-machine-learning/](https://blogs.microsoft.com/ai/azure-responsible-machine-learning/).
- 950 [77] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes  
951 decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–  
952 215.
- 953 [78] Rudin, C. and Radin, J. (2019). Why are we using black box models in AI when we  
954 don’t need to? A lesson from an explainable AI competition. *Harvard Data Science*  
955 *Review*, 1(2).
- 956 [79] Sadjadi, S. O., Kheyrkhan, T., Tong, A., Greenberg, C. S., Reynolds, D. A., Singer,  
957 E., Mason, L. P., and Hernandez-Cordero, J. (2017). The 2016 nist speaker recognition  
958 evaluation. In *Interspeech*, pages 1353–1357.
- 959 [80] Schooler, J. W. and Engstler-Schooler, T. Y. (1990). Verbal overshadowing of visual  
960 memories: Some things are better left unsaid. *Cognitive Psychology*, 22(1):36–71.
- 961 [81] Schooler, J. W., Ohlsson, S., and Brooks, K. (1993). Thoughts Beyond Words:  
962 When Language Overshadows Insight. *Journal of Experimental Psychology: General*,

- 963 122(2):166–183.
- 964 [82] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D.  
965 (2017). Grad-cam: Visual explanations from deep networks via gradient-based localiza-  
966 tion. In *Proceedings of the IEEE International Conference on Computer Vision*, pages  
967 618–626.
- 968 [83] Siau, K. and Wang, W. (2018). Building trust in artificial intelligence, machine learn-  
969 ing, and robotics. *Cutter Business Technology Journal*, 31(2):47–53.
- 970 [84] Slack, D., Hilgard, S., Jia, E., Singh, S., and Lakkaraju, H. (2020). Fooling lime  
971 and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the*  
972 *AAAI/ACM Conference on AI, Ethics, and Society*, AIES ’20, page 180–186, New York,  
973 NY, USA. Association for Computing Machinery.
- 974 [85] Soon, C. S., Brass, M., Heinze, H. J., and Haynes, J. D. (2008). Unconscious deter-  
975 minants of free decisions in the human brain. *Nature Neuroscience*, 11(5):543–545.
- 976 [86] Sporer, S. L., Penrod, S., Read, D., and Cutler, B. (1995). Choosing, Confidence,  
977 and Accuracy: A Meta-Analysis of the Confidence-Accuracy Relation in Eyewitness  
978 Identification Studies. *Psychological Bulletin*, 118(3):315–327.
- 979 [87] ten Brinke, L., Stimson, D., and Carney, D. R. (2014). Some Evidence for Uncon-  
980 scious Lie Detection. *Psychological Science*.
- 981 [88] ten Brinke, L., Vohs, K. D., and Carney, D. R. (2016). Can Ordinary People Detect  
982 Deception After All? *Trends in Cognitive Sciences*, 20(8):579–588.
- 983 [89] The Royal Society (2019). Explainable AI: the basics policy brief-  
984 ing. Available at [https://royalsociety.org/-/media/policy/projects/explainable-ai/  
985 AI-and-interpretability-policy-briefing.pdf](https://royalsociety.org/-/media/policy/projects/explainable-ai/AI-and-interpretability-policy-briefing.pdf).
- 986 [90] Todorov, A. (2017). *Face value: The irresistible influence of first impressions*. Prince-  
987 ton University Press.
- 988 [91] Todorov, A., Mandisodza, A. N., Goren, A., and Hall, C. C. (2005). Inferences  
989 of competence from faces predict election outcomes. *Science (New York, N.Y.)*,  
990 308(5728):1623–6.
- 991 [92] Toreini, E., Aitken, M., Coopamootoo, K., Elliot, K., Gonzalez-Zelaya, C., and van  
992 Moorsel, A. (2020). The relationship between trust in AI and trustworthy machine  
993 learning technologies. In *Conference on Fairness, Accountability, and Transparency*  
994 (*FAT\* ’20*), Barcelona, Spain.
- 995 [93] Towler, A., White, D., and Kemp, R. I. (2017). Evaluating the feature comparison  
996 strategy for forensic face identification. *Journal of Experimental Psychology: Applied*,  
997 23(1):47.
- 998 [94] Tversky, A. and Kahneman, D. (1981). The framing of decisions and the psychology  
999 of choice. *Science*, 211(4481):453–458.
- 1000 [95] Tversky, A. and Shafir, E. (1992). The Disjunction Effect in Choice Under Uncer-  
1001 tainty. *Psychological Science*, 3(5):305–309.
- 1002 [96] Ustun, B., Spangher, A., and Liu, Y. (2019). Actionable Recourse in Linear Classifi-  
1003 cation. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*,  
1004 *FAT\* ’19*, pages 10–19, New York, NY, USA. Association for Computing Machinery.

- 1005 event-place: Atlanta, GA, USA.
- 1006 [97] Wachter, S., Mittelstadt, B., and Russell, C. (2017). Counterfactual explanations  
1007 without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.*,  
1008 31:841.
- 1009 [98] Weller, A. (2019). Transparency: Motivations and challenges. In *Explainable AI:  
1010 Interpreting, Explaining and Visualizing Deep Learning*, pages 23–40. Springer.
- 1011 [99] Wilson, T. D. and Bar-Anan, Y. (2008). The unseen mind. *Science*, 321(5892):1046–  
1012 1047.
- 1013 [100] Wilson, T. D. and Schooler, J. (1991). Thinking too much: Introspection can reduce  
1014 the quality of preferences and decisions. *Journal of Personality and Social Psychology*,  
1015 60(2):181–192.
- 1016 [101] Wixted, J. T., Mickes, L., and Fisher, R. P. (2018). Rethinking the Reliability of  
1017 Eyewitness Memory. *Perspectives on Psychological Science*, 13(3):324–335.
- 1018 [102] Woodruff, A., Fox, S. E., Rousso-Schindler, S., and Warshaw, J. (2018). A qualita-  
1019 tive exploration of perceptions of algorithmic fairness. *Conference on Human Factors  
1020 in Computing Systems - Proceedings*, 2018-April:1–14.
- 1021 [103] Yeung, N. and Summerfield, C. (2012). Metacognition in human decision-making:  
1022 Confidence and error monitoring. *Philosophical Transactions of the Royal Society B:  
1023 Biological Sciences*, 367(1594):1310–1321.
- 1024 [104] Zhao, Q. and Hastie, T. (2019). Causal Interpretations of Black-Box Models. *Journal  
1025 of Business & Economic Statistics*, 0(0):1–10.
- 1026 [105] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. (2016). Learning  
1027 deep features for discriminative localization. In *Proceedings of the IEEE conference on  
1028 computer vision and pattern recognition*, pages 2921–2929.
- 1029 [106] Zhou, X. and Jenkins, R. (2020). Dunning–Kruger effects in face perception. *Cog-  
1030 nition*, 203(January).
- 1031 [107] Zintgraf, L. M., Cohen, T. S., Adel, T., and Welling, M. (2017). Visualizing Deep  
1032 Neural Network Decisions: Prediction Difference Analysis.