Stakes and Understanding the Decisions of Artificial Intelligent Systems

Eva Schmidt

1. Introduction

As artificial intelligent (AI) systems become more and more widespread, their decisions and recommendations increasingly impact people's lives. The predictions of modern AI systems are often more accurate than those of human decision-makers; but this tends to come at the price that they are opaque – not even their developers can understand how the systems work or how they came to particular outputs. This is the so-called black box problem (e.g. Bathaee 2018, Zednik 2019): Where AI systems substantially contribute to high-stakes decisions, it seems especially important that we understand why they provide certain outputs or how they function; but this information is unavailable especially for modern, powerful AI systems – they are black boxes.

Against this backdrop, the last decade has seen an explosion of interdisciplinary work on explainable artificial intelligence (XAI). The point of much of this research has been to alleviate the opacity of AI systems by providing explanations either locally of individual outputs, or globally of a system's overall functioning (REFs XAI overviews). An important philosophical contribution to the field has been to highlight that explanations are useful to the extent that they make AI systems or their output *understandable* to relevant *stakeholders* in the *context* in question (Langer et al. 2021, Nyrup & Robinson 2022, Páez 2019, Fleisher 2022). That is to say, we shouldn't expect that a 'one size fits all' explanation can be made to work for all uses of AI systems. Rather, the context in which a system is employed has an impact on what kinds of explanations and explainability methods will be useful. One crucial element of the context is the intended recipient of the explanation – an explanation succeeds in making an AI system explainable given that the recipient understands the system, or understands why it gave a certain output (Beisbart & Räz 2022). The person's understanding will then contribute to the fulfillment of certain desiderata that are relevant in the context. For instance, it may enable an end user to reasonably trust the system, to bear responsibility for her AI-supported decision, or to detect whether the system's outputs are due to algorithmic bias (Baum et al. 2022, Schmidt 2022).

This line of thought puts into focus the question under what conditions an agent understands something that's explained to her – a question which is the focus of the current paper. In particular, I will investigate how understanding depends on what is at stake in a particular context. Let me motivate this thought. It is often presumed that whether we need explainability at all, or how detailed or accurate explanations have to be, depends on whether we are facing a high-stakes or a low-stakes situation (House of Lords 2018, Adadi & Berrada 2018, Yuan et al. 2023, or the proposed AI Act[1]). For instance, where an AI system determines a listener's music playlist, very limited and superficial explanations of the music picks seem good enough, if any are needed at all; however, in case an AI system calculates the recidivism risk score for a convict, which then influences the severity of the convict's sentence, a good explanation will have to be quite accurate and more detailed. Since the point of XAI is to provide understanding, the following claim is plausible: Whether an explanation suffices to ensure understanding depends on how much is at stake in a concrete situation. In other words, understanding depends on the stakes.

This makes good pragmatic sense: Where AI systems support or take over potentially life-changing decisions, it is important that we can be sure they work flawlessly and that we grasp how they work or how their outputs come about. However, it would be a waste of resources to technically ensure (a deep) understanding of these matters in situations where AI systems make or support clearly harmless decisions. In a more philosophical vein, this claim is highly intuitive, and I will support the intuition with a pair of cases in the following section. I will further use the pair of cases to spell out how exactly the stakes affect understanding and, in particular, understanding why. To do so, I will connect discussions of the concept of understanding (Baumberger et al. 2016) with debates on pragmatic encroachment (Gao forthcoming) and on inductive risk (Hempel 1965, Douglas 2000). My discussion will thereby fill two scientific lacunae: First, little research so far has been done on whether pragmatic factors or non-epistemic values affect understanding (though see Phillips 2020, Kelp 2015, and Wilkenfeld 2013), in addition to knowledge or the rationality of accepting or rejecting scientific hypotheses. Second, while the notion of inductive risk has been applied in several papers to the issue of design of AI algorithms and models (e.g. Johnson forthcoming, Biddle 2022, Karaca 2021),

[1] See https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence.

little attention has been paid to whether the concepts of inductive risk or pragmatic encroachment can be employed to illuminate how practical stakes affect the understanding that is supposed to be generated by explainable AI (though see Sullivan 2022a, 2022b).

The aim of this paper is then to provide a pragmatic encroachment/inductive risk based account of how the stakes affect the understanding of the recipients of XAI explanations. That is to say, I will apply these notions to the understanding of individuals in XAI contexts in order to explain how an individual's understanding, when confronted with an explanation of an AI system's output, can be affected by how much is at stake. In one direction, that notions of pragmatic encroachment and inductive risk can be fruitfully applied in these contexts lends further support to the claim that pragmatic encroachment and inductive risk are genuine phenomena; in the other direction, that there are conceptual tools like pragmatic encroachment and inductive risk, which have been developed for other contexts, but are a great match here as well, lends additional plausibility to the claim that what's at stake in XAI contexts indeed affects understanding.[2]

The plan for this chapter is the following: In section 2, I introduce a pair of low stakes/high stakes cases and, on their basis, tease out the intuition that understanding why an AI system provided a certain output depends not just on the given explanation, but also on how much is at stake. To lay the foundation for my account of how the stakes affect understanding, I next provide a brief overview over some core features of understanding why, and corresponding necessary conditions, that I take from recent philosophical debate (section 3). I then turn to two ways in which the stakes make a difference to whether an agent understands why an output was provided: The stakes affect how good the subject's *evidence* has to be for understanding (section 4) and they affect how well the subject's beliefs about why the system provided its output have to *cohere* internally or with her background beliefs for understanding (section 5). Section 6 concludes.

2. Two Cases of XAI

Here is the pair of cases. In each case, a subject wants to understand why an AI system provides a certain output. Imagine that each system has an equally reliable, built-in function for generating an explanation of a particular output upon demand, and that the subjects in both cases request and receive such an explanation, and that the explanation is correct. Further imagine

[2] Beyond this, the purpose of this chapter is *not* to give a principled defense of pragmatic encroachment or inductive risk on the one hand, or of effects of the stakes, in XAI contexts, on whether an explanation suffices for understanding. It is to propose an account of how exactly the stakes may affect understanding, granted that they do so.

that the subjects in both cases have little to no prior background knowledge of how AI systems work, or of how their particular system works.

> *Playlist* (low stakes)*:* Lizzy exclusively listens to country pop on Spotify. Suddenly, Spotify plays *Ace of Spades* by Motörhead, a song Lizzy strongly dislikes. She wants to find out why this song was played and receives the explanation: “Spotify played *Ace of Spades* because people with a political orientation similar to yours enjoyed listening to this song.”

> *Sentencing* (high stakes): Juana is a judge who has to decide on the sentence length for a Black convict, Connor, and is supported by COMPAS in her decision-making (Northpointe 2015, Angwin et al. 2016). COMPAS assigns a maximum recidivism risk score of 10 to Connor. Upon reviewing the case files, Juana estimated Connor’s recidivism risk as rather low. In light of this, she wants to understand why COMPAS has such a negative prognosis for him. She receives the explanation: “COMPAS assigned a risk score of 10 to Connor because convicts with music interests similar to Connor’s often committed further crimes within two years.”

Intuitively, in *Playlist*, Lizzy understands why Spotify plays *Ace of Spades*, whereas in *Sentencing*, Juana does not understand why COMPAS provides the negative assessment of Connor’s recidivism risk. For instance, it would be absolutely reasonable for Lizzy to think: “Weird, but ok – apparently many others who, like me, don’t care about politics one way or another enjoy listening to Motörhead, and that’s why Spotify played the song, so I get it.” But it would be perfectly reasonable for Juana to think: “Why would COMPAS give Connor such a high risk score because of his *music interests*? I still don’t understand why it gave this prognosis!” At the same time, it would be unfitting for Juana to think: “Weird, but ok – apparently many convicts with music interests like Connor’s have recidivated, and that’s why COMPAS outputted the high risk score, so I get it.”

Let me elaborate some more on these cases before placing them in the context of the pragmatic encroachment debate. Both of them depart from reality in several respects: First, neither Spotify nor COMPAS actually makes available the option for subjects to receive automated explanations of its outputs. However, there are digital tools that automatically provide explanations – for instance, Google search allows users to get an explanation of why specific search results are presented, so such a service could plausibly be made available for these systems as well. There is also so far no consensus on how to evaluate the reliability of an explainability method, though proposals are discussed in the literature (e.g. Nauta et al. 2023, Nielsen et al. 2022).

Second, the explanations provided to the subjects in both cases bring up features that are not actually available to the systems. For example, the questionnaire on which COMPAS relies as input has no entry for music interests.[3] I rely on these concrete example explanations because neither makes any sense (to the subjects or, presumably, to my readers): It is not immediately clear, nor easy to figure out, why political orientation should predict a preference for Motörhead, or why music interests should predict recidivism.[4] At the same time, both explanations appeal to real correlations that AI systems might use for their predictions. Studies have shown a correlation between political orientation and music preference, where listeners who were politically neutral liked listening to heavy metal (Peterson and Christenson 1987), and such features could be used to predict what songs listeners will want to hear (Laplante 2014). As to the *Sentencing* case, music interests or preferences predict race. For instance, rap music has a greater fan base among Black listeners. Say that the system's output is due to a racial bias – it might then use music interest as a proxy attribute for race (Marshall and Naumann 2018, Tschantz 2022; regarding algorithmic bias, see Fazelpour and Danks 2021).

The described cases are supposed to parallel the so-called 'bank cases' that are standard fare in the debate over pragmatic encroachment on knowledge (see e.g. Stanley 2005, Fantl and McGrath 2002, Schroeder 2012). A subject is considering whether to deposit her check in the bank immediately despite the long lines in front of the bank. She has solid, but not perfect, evidence that the bank will be open tomorrow, on Saturday. In the low stakes version of the case, it is of no importance whether the money is deposited quickly or not, and here it's intuitive that the subject knows that the bank will be open tomorrow. In the high stakes version of the case, it is of great importance that the money is deposited before Sunday, since a large mortgage payment will be taken out of the subject's bank account then. In this version of the case, the

[3] See https://www.documentcloud.org/documents/2702103-Sample-Risk-Assessment-COMPAS-CORE.

[4] That said, if readers prefer cases involving more realistic explanations, they can imagine, for example, that Spotify gives the explanation: "Spotify played *Ace of Spades* because listeners in your geographic region enjoyed listening to this song.", and COMPAS: "COMPAS assigned a risk score of 10 because convicts with the same ZIP code as Connor often committed further crimes within two years." Here it is important to keep in mind that both subjects are unable to make sense of these explanations, as they lack the background knowledge to connect them to the given output.

subject intuitively doesn't know that the bank will be open tomorrow, exactly because the stakes are high – if the subject falsely believes that the bank will be open tomorrow and acts on this belief, this will have extremely negative practical consequences for her. Crucial characteristics of these cases are that, first, they are epistemically identical – the subject has the same evidence, the same intellectual capacities and background knowledge – and, second, the only pertinent difference is how much is at stake in each case.

Playlist and *Sentencing* share these characteristics. Recall that I described the explanation-providing functions of the systems as equally reliable – say that, in 90% of cases, the explanations generated by Spotify and by COMPAS highlight central factors bearing on their outputs. So the explanations provided are both equally good evidence. (In *Playlist*, Spotify's explanation is evidence that it played *Ace of Spades* because of Lizzy's political orientation, and in *Sentencing*, the explanation provided by COMPAS is evidence that it ascribed Connor a high risk score because of his musical interests.) Both are true explanations. Neither explanation makes sense from the point of view of its recipient. Political orientation seems completely unconnected to whether one likes a certain music genre; more specifically, that one is politically disengaged seems unconnected to liking Motörhead. Music interest seems completely irrelevant to recidivism probabilities, including to whether a convict is especially likely to recidivate. Finally, Lizzy and Juana both have no relevant background knowledge and equal intellectual capacities.

Both cases differ with respect to the stakes, with respect to how much practically hangs on the subject getting the explanation right. *Playlist* is a low-stakes scenario. If Lizzy were to accept the explanation and take herself to understand why Spotify played *Ace of Spades*, but the explanation were incorrect or misleading, for instance because it leaves out relevant difference-makers, this would have no negative practical consequences. It doesn't really matter whether the song was played because of Lizzy's political orientation, or if there is another better, more pertinent explanation. By contrast, *Sentencing* is a high-stakes scenario, and a lot hangs on Juana getting the explanation right. If Juana were to accept the provided explanation and take herself to understand why COMPAS assesses Connor as high risk, even though the explanation was misleading – in the given scenario, it misleads by omitting the fact that algorithmic bias was a crucial factor – this would have very bad practical consequences. For if she were thus to get wrong core factors that really stood behind the system's output, she would very likely discriminate against Connor by giving him a severe sentence because of algorithmic bias.

Notice the difference between how pragmatic factors affect understanding in these cases and pragmatic encroachment on knowledge: There, the pragmatic encroaches on knowledge just in case the subject's having a false belief and acting on it would lead to very bad consequences (see Schmidt 2023). For understanding, the issue is: If the subject got the actual explanation of the system's output wrong and acted on this basis, would that lead to very bad consequences? 'Getting the explanation wrong' covers scenarios in which the explanation accepted by the subject is false, but also scenarios in which it merely leaves out central contributing factors, so that the subject is *misled* about what really explains why the output was given. These points will become relevant below.[5]

In *Playlist* and *Sentencing*, then the subjects are epistemic equals as far as the quality of their evidence, their intellectual capacities, or their relevant background knowledge go. But they intuitively differ with respect to their understanding of why the relevant output was given: While Lizzy understands why Spotify played *Ace of Spades*, Juana does not understand why COMPAS assigned a high recidivism score to Connor. Since a striking difference between the cases is how much is at stake in the subjects getting the explanations right, it seems that the stakes affect understanding as much as they affect knowledge. To make this claim plausible and to see how exactly the pragmatic encroaches on understanding, we first need to take a closer look at what understanding is.

3. Understanding: The Concept

The mental state of understanding is a currently much discussed topic in the philosophy of science and in epistemology (see Baumberger et al. 2016, Grimm 2021, Hannon 2021 for recent overviews). One topic of debate is the nature of understanding. Similarly to the long-running discussion on the correct analysis of knowledge, there are efforts to provide the conditions under which a subject understands. My aim in this section is to bring out the conditions of understanding that, on my proposal, are affected by high stakes. I will present some core elements of the debate over the nature of understanding, and particularly focus on four widely accepted necessary conditions on understanding.

Let me start with an important background distinction, that between objectual and explanatory understanding (e.g. Baumberger et al. 2016, 5). A person has *objectual* understanding when she understands an overall subject matter or a broader domain, as when she understands machine learning or genetics. We say the person has *explanatory* understanding, or understanding why, when she understands why something happened or is the case, for instance, when she

[5] Similarly, see Biddle and Kukla's (2017) push for broadening our conception of inductive risk to one of epistemic risk, which goes beyond risks from acting on false belief.

understands why inflation went up in 2022 or why the dinosaurs became extinct. Objectual understanding is plausibly relevant to XAI in situations where we want understanding of how an AI system works overall. My interest here is in understanding why a system gave a certain output, which I take to be an instance of explanatory understanding (simply 'understanding' in the following).[6] This is what I will focus on here.

So what are the necessary conditions on (explanatory) understanding? As with conditions on knowledge, there is disagreement about the details, but many views converge on four conditions that a subject has to meet to understand: (1) belief/acceptance, (2) grasp, (3) truth/factivity, and (4) justification (see e.g. Baumberger et al. 2016, Pritchard 2010, chap. 4, Hills 2016, Malfatti 2023). The parallels with the traditional analysis of knowledge, according to which knowledge is justified true belief, are noticeable; a difference is that for knowledge that *p*, the necessary conditions apply only to a single proposition *p*. By contrast, for understanding why *p* is the case, propositions beyond *p* are relevant, viz. those propositions *q*, *r*, *s* that constitute the *explanans* of why *p*, as well as the proposition that *q*, *r*, *s* explain why *p*. With this in mind, let's look at the four necessary conditions on understanding.

S understands why *p* is the case only if

(1) *S believes* or accepts that *p* and that *q*, *r*, *s* (and maybe also believes or accepts that *q*, *r*, *s* explain why *p*),

For instance, I understand why the dinosaurs became extinct 66 million years ago only if I believe that the dinosaurs became extinct then, and I additionally believe propositions such as: that a big asteroid hit Earth, that this caused huge amounts of dust to enter the atmosphere, leading to temperature drops, which led to mass extinction, and that in this way the asteroid impact explains why the dinosaurs went extinct.

Some philosophers relax the belief condition and instead endorse an acceptance condition on understanding. On this variant, *S* does not have to believe the relevant propositions, but may merely accept them as premises from which to draw further inferences, or on which to act. Unlike belief, acceptance is an attitude that a person can correctly have towards propositions that are close to the truth (or approximate it), but are not literally true (Elgin 2017, chap. 2;

[6] Though see Páez (2019, 454) for the view that the understanding involved is still objectual understanding, merely localized.

Malfatti 2023, chap. 8). To illustrate, the dinosaurs didn't go extinct *exactly* 66 million years ago, and not all of them did, seeing as birds are still around. And the explanation sketched above oversimplifies, since many species went extinct immediately due to tsunamis, earthquakes, and wildfires caused by the asteroid impact.[7]

(2) *S grasps* the interconnections between *p* and the propositions *q*, *r*, *s*, which explain it,

Understanding is more than just knowing or believing individual pieces of information, it involves having a handle on how these connect or fit together. The notion of grasp is supposed to capture this. In the example, I understand why the dinosaurs became extinct only if I grasp how their extinction relates, e.g., to the asteroid's impact on Earth, to the dust in the atmosphere, or to the temperature dropping, and how these explanatory factors in turn hang together. More specifically, I have to see what explanatory or counterfactual relationships exist between the dinosaurs' extinction, the asteroid impact, and so on (Riggs 2003). Grasp is often spelled out by way of relevant abilities the subject has to have, such as being able to explain how the extinction happened, or to say what would have happened in certain counterfactual situations, e.g. if the asteroid had bypassed Earth, and so on (Hills 2016).

(3) at least central propositions among *p*, *q*, *r*, *s* (and the proposition that *q*, *r*, *s* explain why *p*) are *true*, or approximately true,

It is highly plausible that I cannot understand why the dinosaurs became extinct if this isn't even approximately true, or if the explanation I believe isn't even in the neighborhood of the truth. Whether the exact truth is needed or something more approximate is sufficient is contentious. Positions range from the view that all the mentioned propositions need to be true (Kelp 2015), to the view that central propositions have to be true, while propositions peripheral to the state of affairs whose understanding is in question may be false (Kvanvig 2009), to the view that the relevant propositions need be merely approximately true to allow for understanding (Elgin 2007). I remain neutral between these positions.

[7] I remain neutral between stricter and more relaxed versions of this condition. For ease of exposition, I will speak of belief and the belief condition in the following.

(4) *S*'s beliefs that *p, q*, *r*, *s* (and that *q*, *r*, *s* explain why *p*) are *justified*: they are (a) supported by *sufficient evidence* and they (b) *cohere* among each other and with *S*'s background beliefs.

If a subject just randomly accepts certain propositions (whether on the *explanandum* or on the *explanans* side), although it's utterly nebulous to her why they should be true, she doesn't thereby gain understanding (Pritchard 2010, 76, referring to Zagzebski 2001). Instead, her beliefs must be justified to constitute understanding. In my view, this has two aspects. On the one hand, the relevant propositions have to be sufficiently justified by the subject's reasons or evidence; on the other hand, they have to provide mutual support to each other by cohering well: by being consistent or standing in explanatory and probabilistic relations with each other (Kvanvig 2003, Elgin 2007, 35) and fitting in well with the other beliefs of the subject (Malfatti 2023, 38, Elgin 2017). For instance, my understanding of why the dinosaurs became extinct presupposes that I don't just randomly believe that an asteroid struck Earth 66 million years ago, but that I have evidence that this is so, e.g. by reading it in a reliable science book. Furthermore, there have to be interconnections between the relevant propositions (which I grasp, see (2)). If I justifiably believe several true propositions in addition to my belief that the dinosaurs became extinct, but they are completely unrelated to each other or to the dinosaurs' extinction, this doesn't amount to understanding. Moreover, imagine that, beyond these beliefs about dinosaurs, the asteroid, and so on, I am a dyed-in-the-wool creationist who believes that God created the world, including ready-made dinosaur fossils, 6000 years ago. Given the lack of fit (or rather, the contradictions) between my beliefs about the dinosaur extinction and my background beliefs, I don't *really* understand how the dinosaurs became extinct.

Some object to a justification condition on understanding and argue that we can understand why something is the case even where the beliefs involved in our understanding are supported only by *defeated* evidence (Hills 2016; Malfatti 2023, 189). To take Malfatti's example, if a medical doctor hears a patient's report of his condition and infers that he suffers from anemia, and then receives the misleading evidence that the patient is a compulsive liar, her evidence (the patient's report) is defeated. But, this line of thought has it, her understanding of why the patient exhibits certain symptoms is not thereby undermined.

Discussing this general point here would lead me too far afield. However, I find this claim implausible at least in XAI contexts, even for low stakes scenarios. Imagine that the advertisements in your Facebook stream seem to pick up on the fact that you just booked a vacation, and that you do not want this. In the Facebook settings, you look up which of your personal information Facebook advertisements use and see that they do *not* rely on your recent purchases. You thereby take yourself to understand something about how your Facebook advertisements are generated, and *trust* Facebook advertisements not to use information about recent purchases. Now imagine that from a reliable source, you receive the misleading defeater that Facebook's statements about which information it uses to personalize advertisements are highly unreliable. This defeats your evidence that Facebook doesn't use information about your booked vacation for its advertisements, so your belief that the ads don't respond to your booked vacation is no longer justified. I submit that at the same time (and in contrast with Malfatti's and others' claims), you also lose your understanding of how Facebook came to display the ads in question to you. I think it would be perfectly reasonable for you to think: "I thought I understood which factors Facebook uses to determine which advertisements to display – but now I'm absolutely clueless. I don't understand why it would show me these specific ads!"

My suggestion is supported by the fact that now you rightly no longer trust Facebook's algorithm not to rely on your recent purchases. Put more generally, understanding can be instrumental to reasonable trust in AI contexts (e.g. Ribeiro et al 2016, EU High-Level Expert Group on Artificial Intelligence 2019, 13, Schmidt 2022). Given this, the fact that your trust is lost indicates that your understanding is lost also. More generally, the call for XAI is not a call for what one might call *idle* understanding, but for understanding that enables stakeholders to *do* things relevant in the context (Langer et al. 2021); but purported understanding based on defeated evidence is insulated from (reasonable) action. At least in XAI contexts, defeaters thus undermine the subject's understanding together with her evidence, and (4) stands as a necessary condition for understanding.

How does all this connect to the claim, introduced in section 2, that the stakes affect whether a subject understands? I will propose in sections 4 and 5 that the stakes affect understanding by making a difference to whether the subject's beliefs are *justified*. The idea to be developed is this: The more is at stake, the stricter the justification condition – both in terms of whether the evidence is strong enough to suffice for justification, and in terms of whether the believed propositions cohere well enough to render the beliefs justified.

4. How the Stakes Affect Understanding (I): Strength of Evidence

Understanding (in XAI contexts) requires epistemic justification of the relevant beliefs, and justification requires sufficient evidence or reasons, just as in the case of knowledge. Correspondingly, one way for the stakes to make a difference to a subject's understanding is by making a difference to whether her evidence or epistemic reasons suffice for justification. How much is at stake in a context can affect understanding via 'standard' pragmatic encroachment.

To illustrate, first imagine a variant on *Playlist* and assume that Lizzy receives the correct explanation that *Ace of Spades* was played because people with a similar listening history enjoyed the song. This is a low-stakes scenario; if Lizzy simply adopted the explanation and acted on it, even if the explanation were false, nothing bad would happen in consequence. Intuitively, her belief – that Spotify played *Ace of Spades* because people with a similar listening history liked the song – is justified by the explanation she receives from Spotify, and she understands why the song was played.

Second, consider the question whether Juana in *Sentencing* understands why COMPAS outputs a maximal risk score for Connor. Imagine (contrary to the original scenario) that the explanation that Juana is given for why COMPAS assigned Connor the high risk score is that individuals with a similar criminal history have often reoffended. Say that Juana knows that algorithmic bias is a real problem of some AI systems, possibly including COMPAS (Angwin et al. 2016, Chouldechova 2017). Finally, imagine that the explanation is correct, and that COMPAS really is sensitive to features of Connor's criminal history shared by convicts who frequently recidivated, and that this explains the system's prognosis.

This is a high-stakes situation: If – counterfactually – an algorithmic bias against Black people really explained the system's prognosis (so the automatically generated explanation of the prognosis was false), this might lead to very bad moral consequences. For if Juana took the provided explanation at face value, formed the false belief that COMPAS provided the high risk score because of Connor's criminal history, and gave him a severe sentence because of COMPAS's high risk score, she would discriminate against him. On the one hand, being imprisoned unnecessarily would be terrible for Connor, and the important moral and legal value of fair treatment in criminal proceedings would be violated. On the other hand, it would be bad for Juana, who we can take to be a conscientious, fair-minded judge, to be put in the position of unwittingly wronging a member of a marginalized group.

In light of this, Juana's belief that COMPAS provided the high risk score because of Connor's criminal history is *not* justified. True, the fact that a reliable explanatory mechanism gave the explanation – that Connor was assigned a maximum risk score because individuals with a similar criminal history have often reoffended – indicates that Connor's criminal history explains COMPAS's high risk score; it is good evidence that COMPAS provided the high risk score because of Connor's criminal history. This evidence justifies belief in this explanation to a degree. However, because so much is at stake, this evidence does not suffice to justify Juana's corresponding belief, all things considered.[8] Think about it this way: Given that Juana is aware of the real-world problem of algorithmic bias, and given that (as in the original *Sentencing*) her own perusal of Connor's files led her to asses him as low risk, is it epistemically rational for her to simply accept the explanation provided by COMPAS? I think not; rather, it would be rational for her to remain in doubt as to the real explanation of the high risk score. As I have phrased it elsewhere (Schmidt 2023), where a lot hinges on an agent's getting a belief right, she needs to be especially *diligent* in weighing her reasons in forming the belief, which is to say that it takes more for her epistemic reasons to believe to suffice for all things considered justified belief.[9] In the current context, this means that Juana needs to be especially sensitive to possible limitations of her evidence, such as the possibility that the explanation she received might be incorrect or misleading. She should therefore accord the testimony provided by COMPAS's explanatory mechanism less weight than she would if nothing much was at stake. Given that her evidence is thus weakened, it is insufficient to justify outright belief; and so her understanding of why COMPAS provided its output is undermined.[10]

The same point can be made with the help of the notion of inductive risk, which is a controversial topic in the philosophy of science (Hempel 1965, Douglas 2000, Steel 2010):[11] There is always some probability of error in accepting or rejecting a scientific hypothesis on the basis of empirical evidence. Errors of two kinds are possible: First, false positive error – the hypothesis that a certain phenomenon occurs is accepted, although it does not occur; second, false negative error – the hypothesis that a phenomenon occurs is rejected, but it does occur.

[8] I rely on the distinction between a graded notion and an all things considered notion of justification: A proposition can be more or less justified for a believer, where this does not fix whether this justification suffices to make outright belief right. Contrasting with this, there is a kind of on-off justification (either a belief has it fully, or it doesn't have it at all) that we can think of as knowledge-level justification – it is this kind of justification we have in mind when we take justification to be one of the necessary conditions for knowledge (Schroeder 2012).

[9] My full account of pragmatic encroachment in Schmidt (2023) spells this thought out by appeal to *attenuating conditions* that weaken the pertinent reasons to believe. It further appeals to how much weight a *perfectly virtuous reasoner* or *phronimos* would accord her reasons, and uses this to explain why the subject's reasons are weakened. There is no need to take these claims on board to accept my proposal here. Note, however, that this dovetails nicely with Biddle and Kukla's (2017, 217) account of "phronetic risks", to be briefly discussed in the following section.

[10] Another respect in which the evidence provided by COMPAS may be insufficient is that it is bare statistical evidence, where it is important to make judicial decisions on the basis of individualized evidence (Schmidt et al. 2023, Sullivan 2022a, 314)

[11] See Miller (2014), Fantl and McGrath (2011) for the relation between pragmatic encroachment and inductive risk.

The risk that one of these two errors will occur is called inductive risk. According to Hempel (1965), scientists cannot avoid this risk; the only latitude they have concerns whether to err on the false positive side or on the false negative side. In other words, there is a tradeoff – scientists can only minimize the risk of false-negative errors at the cost of an increased risk of false-positive results, and *vice versa*. They must thus independently determine how much evidence is needed before the hypothesis can be accepted. To do so, they need to look to *non-epistemic values*. For it is what is of practical or moral value to us that determines which type of error we should rather risk – missing out on true hypotheses or accepting hypotheses that are false (see Douglas 2000, 561/562).

Applied to *Sentencing*, this means that, since false positive errors (mistakenly accepting an automatically generated explanation of why COMPAS deemed Connor high risk which sounds unbiased) raise the risk of violating important moral values, such as fair treatment in criminal procedures, the risk of false positive error should be minimized. Juana shouldn't accept the explanation on the given evidence, but needs better evidence to do so. In other words, the belief that Connor was accorded a maximum risk score by COMPAS because of his criminal history is not justified for her. Since justification is a necessary condition for understanding, she therefore does not understand why COMPAS assessed Connor's recidivism risk so negatively. By contrast, in *Playlist*, no practical or moral values are affected if Lizzy believes Spotify's explanation of why it played *Ace of Spades*, i.e. if she makes a false positive error. There is no need to minimize risking such errors, and so she is justified to believe the explanation, and thereby understands why Spotify played the song.

In the philosophy of science, inductive risk is used by opponents of the so-called 'value-free ideal of science' to argue that not only what is of epistemic value is directly relevant to science, but that matters of practical, moral, or political value affect which theories scientists correctly accept (e.g. Longino 1990, Rudner 1953). Again, the idea is that in many cases, practical or moral concerns set the standards within which scientists justifiably accept hypotheses, choose models, and so on. In recent years, several philosophers have transferred this idea to AI, arguing that the design of AI systems is not, and cannot be, value-free, but that the correct development of a system – e.g. how the model is designed, which data are chosen, or how they are labeled – is partly determined by what is practically or morally at stake (Sullivan 2022a, Sullivan 2022b, Karaca 2021, Johnson forthcoming, Biddle 2022, Ratti & Graves 2022). Sullivan (2022a, 2022b) extends this analysis to how the practical or moral values impact individuals' understanding of AI systems, and so to my topic here. In line with this, my proposal is that there is no suitable value-free account of when explanations suffice for understanding why certain AI outputs were given; any such account must be sensitive to the practical or moral values at stake in the context.

That said, standard pragmatic encroachment, affecting the sufficiency of a subject's evidence, and paralleling this, inductive risk as sketched in the previous paragraphs, give a rather limited picture of how the stakes impact understanding *specifically.* To provide a fuller picture, I now turn to a second way in which the pragmatic encroaches on understanding why. I propose that pragmatic factors also impact justification (as a necessary condition on understanding) via its *coherence* component.

5. How the Stakes Affect Understanding (II): Coherence

My account as developed so far does not concern any aspects specific to understanding, but proposes that practical or moral values affect understanding exactly in the same way as they affect knowledge: by bearing on how much evidence is needed for all things considered justification. So the account so far is maybe rather unsurprising. For instance, philosophers who take understanding to be a kind of knowledge will think this exactly the result they expected. However, there is an additional way in which the pragmatic encroaches on understanding which goes beyond this standard story. My proposal is that understanding is harder to acquire in high-stakes contexts because of higher demands on the *coherence* of the provided explanation with the explanandum as well as with the subject's background beliefs, which again makes it harder to meet the justification condition.

To illustrate this additional way in which practical or moral matters affect whether a subject understands, let me contrast again (the original versions of) *Playlist* and *Sentencing*. First, let's ask whether in *Playlist*, Lizzy is justified to believe that Spotify plays *Ace of Spades* because people with a political orientation similar to Lizzy's liked the song. Focus on how well this explanation of why the song was played coheres with her background beliefs about musical beliefs and political orientation: As discussed above, Lizzy is not at all aware that there might be a connection here, either between musical tastes and political views in general, or between being politically neutral and enjoying heavy metal. The explanation doesn't fit well into her pre-existing beliefs about political views or about music preferences; nor does the explanation "because of your political views" cohere well with what Lizzy seeks to explain, why *Ace of Spades* was played. What little coherence there might be comes from the fact that technical devices sometimes do things for odd reasons; also, there is at least no inconsistency. However, there is nothing at stake in Lizzy's believing the explanation, and in her believing that the fact that others with a similar political orientation liked the song explains why Spotify played it. If she were to believe it, and it were misleading or false, there would be no bad consequences. In line with the intuitive judgment I outlined above in sect. 2, that Lizzy does understand why *Ace of Spades* was played in virtue of the explanation, my proposal predicts that her belief is justified and so meets this necessary condition for understanding. In a low-stakes scenario little coherence suffices for justification.

Compare this with *Sentencing*. Recall that the explanation given for COMPAS's negative assessment for Connor – "individuals with similar music interests as Connor often reoffended" – also coheres little with Juana's background beliefs, or with what she is trying to understand. As far as she knows, there is no explanatory or probabilistic relationship between a convict's musical interests and risk of recidivism generally speaking, and *a fortiori* none specifically between Connor's music preferences and his (allegedly) being at high risk of recidivating. So, again, there is very little coherence – although Juana may think that technical devices sometimes to things for odd reasons, and she will see no inconsistency.

In contrast to *Playlist*, in this scenario there is a lot at stake: If Juana were to accept the explanation and take herself to understand why COMPAS assigned Connor a maximum risk score, but the explanation was misleading (by leaving out the important factors that music interest are used by COMPAS as a proxy attribute for race and that the system has a bias against Black people), then Juana's sentencing decision would likely discriminate against Connor (see above). Since a lot is at stake, there are more severe demands on the coherence of the propositions relevant to the subject's understanding, internally and with Juana's background beliefs. The little coherence of the explanation provided by COMPAS for its output with Juana's other beliefs and with what she is trying to understand does not suffice for justification. Since justification is necessary for understanding, and in line with the intuitive judgment teased out in sec. 2, she does not understand why COMPAS assesses Connor so negatively.

The situation for Juana is different if the explanation provided coheres better. Imagine that Juana is not as uninformed as in the original version of the scenario, but that she has the following background knowledge: She is aware that music interests correlate with race in the United States and that some AI systems are affected by algorithmic bias and use proxy attributes for race. Further, imagine that the explanation is correct and that COMPAS gives its output in response to a proxy attribute for race due to algorithmic bias. If so, the explanation is highly coherent for her – she can infer that COMPAS is probably using music interest as a proxy attribute for race and is biased against Connor as a Black convict, and that COMPAS gives

Connor a maximum risk score because of algorithmic bias. In this case, the coherence is high enough to ensure justification and to allow Juana to understand despite the high stakes.

My proposal is distinct from, but relates in interesting ways to that of Sullivan (2022a). Sullivan applies the concept of pragmatic encroachment to high stakes XAI contexts, but appeals to condition (2) on understanding, i.e. to grasp and in particular to the related abilities. She claims that "full-fledged understanding … depends on the threshold of the number of what-if questions in the set of all possible what-if[-things-had-been-different] questions on a given topic in a particular context that is necessary to attribute understanding" (p. 316). The core of Sullivan's proposal is that, depending on how much is at stake, a subject has to be able to answer fewer or more questions about what would happen in counterfactual situations, to count as having outright understanding. Such abilities show the subject's *grasp* of the relations between the relevant propositions (see sec. 3).

Although Sullivan appeals to a different condition on understanding, her view picks up on related features of the subject and her situation: To have counterfactual reasoning abilities (i.e. abilities to answer relevant what-if questions) is, at bottom, to grasp how the different elements of the explanation, the state of affairs to be explained, and surrounding matters cohere. For example, in *Playlist*, Lizzy can give some limited answers to questions like "What would Spotify have played if you had a different political orientation?"; this corresponds to the limited coherence of the explanation provided by Spotify with her background beliefs and with the fact in need of explanation, that Spotify played *Ace of Spades*. By contrast, imagine a scenario in which Spotify plays a new country pop song for Lizzy, and then outputs the explanation that people with a similar listening history enjoyed this song. This explanation coheres extremely well with her background beliefs about new songs that people who enjoy country pop will like. And correspondingly, Lizzy can provide more in-depth answers to the question, "What songs might Spotify have played if you had a different listening history?"

Still, my proposal and Sullivan's are not identical. We can conceive of subjects who struggle to grasp relations between propositions involved in an explanation, and who struggle to answer what-if questions even if they grasp some such relations. In my example cases it's the missing coherence between the given propositions that is problematic for the subjects' understanding, at least given high-stakes, not that they fail to grasp existing coherence relations or answer what-if questions on the basis of grasping them. Both necessary conditions are distinct, as there might be cases where subjects have beliefs that cohere in the right way, but fail on the grasp/abilities dimension. *Prima facie*, both the grasp condition and the justification condition are stakes-sensitive. Sullivan's and my proposals can thus be seen as complementary.

My proposal, with its claim that practical or moral values can impact understanding by making it harder to achieve sufficient coherence, goes beyond inductive risk narrowly conceived. But it is in line with Biddle and Kukla's (2017) broader notion of *phronetic risk* (mentioned in footnote XYZ above). They conceive of phronetic risk as

> epistemic risks that arise during the course of activities that are preconditions for or parts of empirical (inductive or abductive) reasoning, insofar as these are risks that need to be managed and balanced in light of values and interests. (p. 220, italics omitted)

My claim that matters of practical or moral value impact whether a subject's epistemic position is good enough for knowledge or understanding is in the same spirit. Understanding is more difficult to achieve not just where there would be very bad consequences in case the subject acted on a false belief, but also where important practical values might be affected in case she acted on a misleading (even if literally true) explanation, as outlined above.

Let me address a final issue. Some may think that, instead of treating understanding as an on-off phenomenon (one either has it or one doesn't), we should treat it as graded – a subject can have more or less of it (see Pedersen Phillips 2020, 142).[12] Kelp (2016), for instance, develops a graded notion of understanding that takes maximal understanding as its base point; he then defines degrees of understanding by their distance from the maximum. At the same time, however, he introduces a context-relative notion of outright understanding (Kelp 2015, 3813). What degree of graded understanding suffices for outright understanding depends on the practical task at hand, according to Kelp. Although there are some differences between Kelp's account and what I am proposing here, I am happy to allow that a graded concept and an outright concept of understanding should both be acknowledged. I only insist that what is of interest when it comes to the stakes bearing on whether an XAI method suffices to provide understanding in a context is *outright* understanding (and this fits quite well with Kelp's claims). This outright understanding, I submit, works similarly to knowledge: Whether it is achieved depends on whether the relevant beliefs are sufficiently justified, which in turn depends on the stakes.

The following picture of pragmatic encroachment specifically on understanding in XAI contexts has emerged: In low-stakes situations, when a subject's belief that this-and-that explains an AI system's output coheres only weakly, internally or with her background beliefs,

[12] I model the distinction on one from Sturgeon (2008), who uses the "on-off" terminology to distinguish a conception of outright belief from one of graded belief.

this can be good enough for justification on the coherence dimension, so that the justification condition on understanding can quite easily be met. By contrast, when much is at stake, the requirements for how well the relevant beliefs have to cohere are much higher, so that it is harder for the subject's beliefs to be justified. And accordingly, in cases with weak coherence, the relevant beliefs are often not justified, which is to say that the subject does not understand why the AI system gave the output it did.

6. Concluding Remarks

I have started from the assumption, widely accepted by theorists working in the field of XAI, that there are higher demands on explainability in contexts in which important decisions are made or supported by AI systems. This raised the question how exactly the stakes in a context might affect whether an explanation is successful, which is to say, how they affect whether a subject receiving the explanation in the context understands what is explained. To answer this question, I focused on explanations of why a particular output was provided, and looked to philosophical debates on pragmatic encroachment and inductive risk.

According to my proposal, it is the justification condition on understanding that is sensitive to what is at stake in a situation and thus leads to the stakes-sensitivity of understanding – when the stakes are higher, there are more stringent requirements on all things considered justification and thus on outright understanding. I spelled this idea out, first, by appeal to the claim that in high-stakes contexts, the subject must have better evidence or reasons in order to have justification, and second, by appeal to the claim that in such contexts, there must be greater coherence between the propositions involved in the believed explanation, as well as coherence of these with the subject's background beliefs, so that the beliefs pertinent to her understanding can be justified.

My proposal leaves open many questions about how the context bears on genuinely explainable AI. For instance, how do factors such as time pressure or the subject's background understanding and background knowledge of computers or AI systems impact whether an explanation succeeds in making a system explainable to her? And how do these factors interact with the stakes? Answering these questions will have to be left for future research.

References

Adadi, A. & Berrada, M. (2018). *Peeking inside the black-box: A survey on Explainable Artifcial Intelligence (XAI*). IEEE Access 6 (2018), 52138– 52160.

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

Baumberger, C.; Beisbart, C. & Brun, G. (2017). *What is Understanding? An Overview of Recent Debates in Epistemology and Philosophy of Science*. In: S. Grimm, C. Baumberger & S. Ammon (eds.). (2017). Explaining Understanding: New Perspectives from Epistemolgy and Philosophy of Science. Oxford: Routledge, 1-34.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data* 5(2), 153-163.

EU High-Level Expert Group on Artificial Intelligence (2019). *Ethics guidelines for trustworthy AI*. URL: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

Fantl, J., & McGrath, M. (2002). Evidence, pragmatics, and justification. *The Philosophical Review* 111, 67–94.

Fantl, J. & McGrath, M. (2011). *Pragmatic Encroachment*. In: S. Bernecker & D. Pritchard (eds.). (2011). The Routledge Companion to Epistemology (1st ed.). Oxford: Routledge, 558-568. https://doi.org/10.4324/9780203839065

Fazelpour, S. & Danks, D. (2021). Algorithmic bias: Senses, sources, solutions. *Philosophy Compass* 16(8), e12760. https://doi.org/10.1111/phc3.12760

Fleisher, W. (2022). Understanding, Idealization, and Explainable AI. *Episteme,* 19(4), 534-560. https://doi:10.1017/epi.2022.39

Grimm, S. (2021). "Understanding". In: Edward, N. (ed.) (2021). *The Stanford Encyclopedia of Philosophy* (Summer 2021 Edition). URL: <https://plato.stanford.edu/archives/sum2021/entries/understanding/>.

Hannon, M. (2021). Recent Work in the Epistemology of Understanding. *American Philosophical Quarterly* 58(3), 269-290. URL: https://scholarlypublishingcollective.org/uip/apq/article/58/3/269/283447/RECENT-WORK-IN-THE-EPISTEMOLOGY-OF-UNDERSTANDING

House of Lords, Select Committee on Artificial Intelligence. (2018). *AI in the UK: ready, willing and able*? London: House of Lords.

Kelp, C. (2015). Understanding Phenomena. *Synthese* 192(12), 3799-3816.

Kelp, C. (2016). Towards a Knowledge-Based Account of Understanding. In: S. Grimm, C. Baumberger & S. Ammon (eds.) (2016). *Explaining Understanding*. Oxford: Routledge, 251-271.

Laplante, A. (2014). *Improving Music Recommender Systems: What Can We Learn from Research on Music Tastes*. Taipei: International Society for Music Information Retrieval.

Marshall, S. R. & Naumann, L. P. (2018). What's your favorite music? Music preferences cue racial identity. *Journal of Research in Personality* 76, 74-91.

Miller, B. (2014). Science, values, and pragmatic encroachment on knowledge. *European Journal for Philosophy of Science* 4(2), 253-270.

Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., van Keulen, M., & Seifert, C. (2023). From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. ACM *Comput. Surv*. 55(13s), Article 295. https://doi.org/10.1145/3583558

Nielsen, I., Dera, D., Rasool, G., Ramachandran, R. & Bouaynaya, N. (2022). Robust Explainability: A tutorial on gradient-based attribution methods for deep neural networks. *IEEE Signal Processing Magazine* 39, 73-84. https://doi.org/10.1109/MSP.2022.3142719

Peterson, J. B. & Christenson, P. G. (1987). Political orientation and music preference in the 1980s. *Popular Music and Society* 11(4), 1-17.

Phillips, K. P. (2020). *Knowing What Is Said & Other Essays on Knowledge, Understanding, and Communication*. PhD Dissertation.

Pritchard, D. (2010). 'Understanding', *The Nature and Value of Knowledge: Three Investigations.* Oxford Academic. https://doi.org/10.1093/acprof:oso/9780199586264.003.0004.

Ribeiro, M. T., Singh, S. & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In: Association for Computing Machinery (ed.) (2016). *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135-1144.

Riggs, W. D. (2003). Balancing our epistemic goals. *Noûs* 37(2), 342-352.

Schroeder, M. (2012a). Stakes, withholding, and pragmatic encroachment on knowledge. *Philosophical Studies*, 160, 265–285.

Stanley, J. (2005). *Knowledge and Practical Interests*. Oxford University Press.

Sturgeon, S. (2008). Reason and the Grain of Belief. *Noûs* 42(1), 139-165. http://www.jstor.org/stable/25177157

Sullivan, E. (2022). *How Values Shape the Machine Learning Opacity Problem*. In: Lawler, I., Khalifa, K., & Shech, E. (eds.). (2022). Scientific Understanding and Representation: Modeling in the Physical Sciences (1st ed.). Oxford: Routledge, 306-322.

Sullivan, E. (2022b). Inductive Risk, Understanding, and Opaque Machine Learning Models. *Philosophy of Science* 89(5), 1–13. https://doi:10.1017/psa.2022.62

Tschantz, M. C. (2022). *What is Proxy Discrimination?* In: Association for Computing Machinery (2022). Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22). New York. https://doi.org/10.1145/3531146.3533242

Yuan, C. W., Bi, N., Lin, Y.-F. & Tseng, Y.-H. (2023). *Contextualizing User Perceptions about Biases for Human-Centered Explainable Artificial Intelligence*. In: Association for Computing Machinery (2023). Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23). New York. Article 248, 1–15. https://doi.org/10.1145/3544548.3580945