EXPERTS & AI SYSTEMS, EXPLANATION & TRUST

A COMPARATIVE INVESTIGATION INTO THE FORMATION OF EPISTEMICALLY JUSTIFIED BELIEF IN EXPERT TESTIMONY AND IN THE OUTPUTS OF AI-ENABLED EXPERT SYSTEMS

ELIZABETH ANNE SEGER

Trinity Hall

University of Cambridge

May 2022

This thesis is submitted for the degree of Doctor of Philosophy

DECLARATIONS

- This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the preface and specified in the text.
- No substantial part of my thesis has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text.
- This thesis does not exceed the prescribed word limit for the History and Philosophy of Science Degree Committee

TITLE

Experts & AI systems, explanation & trust: A comparative investigation into the formation of epistemically justified belief in expert testimony and in the outputs of AI-enabled expert systems

AUTHOR

Elizabeth Anne Seger

SUMMARY

The relationship between human experts and those that seek their advice (novices), and between AI-enabled expert systems and users, are *epistemically imbalanced relationships.* An epistemically imbalanced relationship is one in which the information source (expert/AI) occupies an epistemically privileged position relative to the novice/user; she/it can utilize capacities, resources, and reasoning techniques to draw conclusions that the novice/user would be unable to access, reproduce, or in some cases, comprehend on her own.

The interesting and problematic thing about epistemically imbalanced relationships is that when the epistemically disadvantaged party seeks out expert/AI aid, then in virtue of the novice's epistemically disadvantaged position, she is not well-equipped to independently confirm the expert/AI's response. Consider for example, a physician who outlines a cancer treatment regime to a patient. If the physician were then to try to explain to the patient how she decided on that specific regime (including drug doses, timings, etc.) it is not clear how the explanation would help the patient justify her belief in the physician's claims. If an expert outlines her reasoning in such detail that it provides strong evidence in support of her claim – for instance, such that a series of true premises logically leads to a conclusion – then the novice is unlikely to have the expertise necessary to recognize the evidence as supporting the claim.

Accordingly, the question stands, how can the novice, while remaining a novice, acquire justification for her belief in an expert claim? A similar question can be asked of user-AI interactions: How can an AI user, without becoming an expert in the domain in which the AI system is applied, justify her belief in AI outputs? If an answer can be provided in the expert-novice case, then it would seem that we are at least on our way to acquiring an answer for the AI-user case.

This dissertation serves a dual purpose as it responds to the above questions. The primary purpose is as an investigation into how AI users can acquire a degree of justification for their belief in AI outputs. I pursue this objective by using the epistemically imbalanced novice-expert relationship as a model to help identify key challenges to user appraisal of AI systems. In so doing, the primary objective is achieved while pursuing the dissertation's secondary purpose of addressing standing questions about the justification of novice belief in human expert claims.

The discussions that follow are framed against an overarching conceptual concern about preserving epistemic security in technologically advanced societies. As my colleagues and I have defined it (Seger et al., 2020), an epistemically secure society is one in which information recipients can reliably identify true information or epistemically trustworthy information sources (human or technological). An investigation into how novices and users might make epistemically well-informed decisions about believing experts and AI systems is therefore an investigation into how we might address challenges to epistemic security posed by epistemically imbalanced relationships.

ACKNOWLEDGMENTS

I think it possible that I have benefited from the support of the best supervisor in the history of PhD supervisors. So, first and foremost, I would like to express my tremendous gratitude to my supervisor, Stephen John, who's academic guidance, countless draft revisions, kind encouragement, admirable patience, and good sense of humor have been essential to the completion of this dissertation and to a genuinely enjoyable PhD experience.

I would also like to thank Shahar Avin for his academic mentorship throughout my Master's and PhD, for presenting me with the opportunity to lead the CSER/LCFI Epistemic Security project, and for supporting me through the learning process of coordinating a large multi-author report. My work on the Epistemic Security project quickly became integrated with my PhD research, and the experience has yielded many exciting opportunities for future work and research. On a similar note, I am grateful to the CSER/LCFI research community for making me feel at home and valued as an adopted member. I would like to extend a special thanks to the following researchers for taking special interest in my work and involving me in theirs: Seán Ó hÉigeartaigh, Stephen Cave, Rune Nyrup, Kanta Dihal, and 2018 summer visitor, Bennett Holman.

A special thanks, also, to my once Master's dissertation supervisor and now friend, Agnes Bolinska, for encouraging me to take the leap in applying for a PhD, and for pushing me to join a rowing team. Mornings on the river have been a saving grace. Similar gratitude is also owed to my secondary PhD advisor, Jacob Stegenga, for encouraging me to apply to the Cambridge HPS PhD, and to Soraya de Chadarevian for recommending that I apply for the HPS Master's program in the first instance. It was only supposed to be one year in the UK! I wonder if you knew what you were starting… I like to think you did.

Of course, none of this would have been possible without the love and support (and patience) of my incredible friends and family. Rebecca Bennett and Alice Fairnie, I would not have survived without you. Josh Hope-Collins, you have a singular ability to peel me off the ceiling and bring me back down to earth. You are amazing, and I love you. My siblings, Julia and Christian, you are, and always will be, my best friends no matter how far apart we live. Grandma Helen, Grandma Maryellen, Grandpa Channey, Grammie Cheryl, and Grandpa George (Gramps), though we do not speak as often as we should, I feel your love every day.

Finally, a massive thank you to Mom for exuding enough warmth and love to leap oceans (and for your meticulous copyediting). Thank you to Steve for teaching me from an early age, that I can do anything I put my mind to, and to Dad for being my rock, always reliable, always there, always ready to listen.

To my family,
given and chosen

CONTENTS

PART 2 TRUST

Introduction

The vast majority of what you think you know about the world, you learned from someone else. Is the world flat or spherical? How hot is the sun? Are vaccines safe? Are they effective? Who was responsible for 9/11? Some things we learn through direct observation (e.g. I know it is raining right now because I see it is raining outside) or by reasoning from observation or memorized facts (e.g. I know the aluminum foil I just took out of the oven will not burn my hand because aluminum does not retain heat well). But mostly, people acquire information about the world by listening to others, by consulting books, searching the web, attending lectures, and asking question of those who either by luck or by training have privileged access to information, such as a witness uniquely positioned to observe a crime or an expert, say a physician, with the specialized knowledge necessary to make a diagnosis. When one has sufficient justification for her belief in claims made by others, she might be said to hold knowledge based on that claim.[1] Knowledge acquired from other people's claims is generally referred to as knowledge from testimony (Coady, 1992; Lackey, 2008; Gelfert, 2014).

Compared to other basic sources of knowledge such as knowledge from perception, memory, or reason, knowledge from testimony is the most prevalent and arguably the most important knowledge source; without the ability to acquire knowledge from others, each new generation would have to rediscover facts known to previous generations making it impossible for humanity to

[1] According to the traditional Justified True Belief (JTB) theory of knowledge, knowledge obtains when a belief is true, and a person holds some adequate threshold of justification for holding that belief (Ayer, 1956:34; Chisholm, 1957: 16). There are, of course, variations on this theory regarding what constitutes adequate justification, how justification can be obtained, and whether justification, truth, and belief are sufficient conditions for knowledge to obtain (See Gettier's (1963) classic rejection of the JTB theory of knowledge), though more detail is not necessary at this time.

accumulate any significant body of knowledge (Fricker, 2006). More so, testimony is essential to executing any large-scale projects that require the collaboration of multiple individuals and the input of multiple areas of expertise (Kukla, 2012; Kitcher, 1990). For example, no one person holds all of the knowledge and skill necessary to design, build, and successfully deploy a rocket to Mars. The project requires the input of experts ranging from physicists and mathematicians to mechanical engineers, material scientists, and telemetry specialists. To collaborate effectively, these experts must be willing to believe each other's contributions where those contributions extend beyond one's own expertise. For instance, if a group of material scientists claim that a new polymer can withstand extremely high heat, the mechanical engineer who might wish to use the polymer in her design of a reentry vehicle will have to take the material scientists at their word unless she becomes a material scientist in her own right. Similarly, responding to the COVID-19 pandemic has required the elicitation of expert input from a variety of fields ranging from epidemiology, virology, and public health, to statistics, economics, and public communications. In this case not only do the experts build upon each other's work, but members of the public also base their decisions about how to act – e.g. whether to wear a mask or receive a vaccine – on their belief in expert claims about the efficacy and/or safety of those activities.

One would be right to point out, however, that most national responses to COVID-19 are not shining examples of how the exchange of testimonial knowledge among experts and public actors enables efficient and effective collaborative efforts. Throughout the pandemic, the marginalization of important information sources has degraded trust in health authorities and slowed public response to the crisis (World Health Organization, 2020). Consider, for instance, the Trump administration's efforts to financially hobble the CDC and to undermine the agency's public health guidance (Reviving the US CDC, 2020). Furthermore, the dissemination of conflicting information about alternative treatment options, vaccine safety, the origins of the COVID-19 virus, and the efficacy of face coverings has increased the difficulty of coordinating unified public response (Hubert et al., 2020; Jourová, 2020).

In a recent report I published with colleagues at the Centre for the Study of Existential Risk (CSER), the Alan Turning Institute, and the Defense Science and Technology Laboratory (Dstl) we refer to interferences to the processes by which people acquire and evaluate information from external sources as 'epistemic threats' to a society's 'epistemic security' (Seger et al., 2020). Epistemic threats include blatant censorship efforts, deceptive disinformation campaigns, the formation of information echo chambers, the erosion of trust in expertise, and so on. An epistemically secure society is one that is robust to such threats such that individuals are able to consistently access and identify true and/or epistemically well-founded information. Preserving a society's epistemic security is important because timely decision-making and collective action are easily undermined by disrupting the processes by which information is gathered and assessed by decision-making bodies and by the public. If there is no shared belief among the actors in a community about the nature of a crisis or the

efficacy of a proposed response, collective action is less likely to ensue. More so, if those beliefs are not epistemically well-founded there is greater risk that collective action based on those beliefs will not have the desired effect. For example, if it is widely, though falsely, believed within a community that public health workers are infecting people with HIV instead of delivering treatment as they claim, then the community will likely resist aid, and the epidemic will continue to spread.[2]

A central aim of our epistemic security report is to identify how various epistemic threats are exacerbated by modern information producing and mediating technologies. Modern information technologies make good decision-guiding information more widely available and easily accessible than at any point in history. This bodes well for epistemic security. However, epistemically well-founded information is not the only information catalyzed by modern information technology. Misinformation is amplified as well. For example, large social media platforms make it easier than ever to spread false information to a wide audience either intentionally or accidentally. More so, mass data processing capabilities allow content to be targeted at "easily convinced" individuals irrespective of its truth value. As such, it is still necessary to sift informational wheat from the chaff, but technologically enabled (mis)information abundance and fine-tuned information targeting capabilities make doing so all the more difficult.

The issue is further complicated in that some of the processes by which people appraise information they receive from others via face-to-face testimony are particularly ill-suited to apprising technologically produced or mediated information. For example, in deciding whether to believe what other people say, people naturally employ quick heuristics like looking for signs that the speaker is sincere and competent – perhaps as indicated by vocal intonation or the dialectical ease with which she produces an explanation for her claims – or by looking to see if many other people agree with the claims being made – we are more likely to believe something, if the claim already seems to be widely accepted (Goldman, 2001; Anderson, 2011). These heuristics for deciding whether to believe a claim are of course fallible, but in general, on the day-to-day, they serve us well. Humans are limited in their cognitive capacity to hold, recall, and manipulate information, so when put on the spot, it is more difficult to fabricate a convincingly detailed lie than to recount what one truly believes. Accordingly, awkward dialog tends to provide some indication of insincerity or incompetence. Furthermore, the larger the number of people who come to believe the same thing by at least partially independent means, then the more likely the belief is to be true.[3]

However, relying on such heuristics can be epistemically detrimental when information is produced or mediated by technological artifacts. For example, in a recent project spearheaded by

[2] Katherine Furman (2016) makes a detailed study of AIDS denialism and misinformation about the spread of HIV and AIDS in early 2000's South Africa.

[3] It is often claimed that beliefs are more likely to be correct if they are held independently by many individuals. For more discussion on this point and what it means to hold independent beliefs, see, among others, Dietrich and Spiekermann (2012), Anderson (2006), Estlund (1994), Lehrer & Wagner (1981), Goldman (2001), and Boland (1989) on knowledge by agreement and Condorcet's Jury Theorem

OpenAI, researchers explore whether human users might be able to judge debates between conflicting AI systems to determine which one they ought to believe (Irving & Askell, 2019; Irving et al., 2018). However, the OpenAI researchers caution that unlike human experts, Artificial Intelligence (AI) systems should not be expected to be dialectically disadvantaged when telling lies or fabricating arguments. Computationally powerful AI systems are not cognitively limited in the same way as humans, and if designed, say to optimize for attracting user trust, it may be more efficient for the system to fabricate explanations to elicit the desired user response. Furthermore, Matt Chessen (2017) describes how AI-enabled automatic content generation technologies like natural language processes systems (NLPs) and audio Deepfakes may be used to identify and mimic speech patterns or vocal tones that users respond to positively in order to gain user trust regardless of the truth-value of the speech content.[4] Consequences of this capability are already being realized. In January 2020, Deepfake audio was used to execute a $35million bank heist by tricking a bank branch manager into initiating an unauthorized money transfer (Anderson, M., 2021).

Chessen also warns that these same technologies can be used to fake widespread subscription to harmful ideas or false beliefs (e.g. vaccines contain toxic doses of mercury). Web robots, called "bots", employ natural language content generation technologies to produce text (headlines, articles, comments) that can be indistinguishable from content produced by humans. A single person with minimal technical knowledge can manage fake profiles for hundreds of bots on social media platforms like Facebook and Twitter where the bots disseminate large volumes of original content to promote the desired idea. In this way, bot deployment presents the appearance of widespread belief in the promoted message which in turn encourages more people to take the message seriously. Furthermore, social media platforms utilize information targeting technologies to primarily expose people to information that is likely to appeal to them as indicated by content the person has consumed in the past. This leads to the formation of "filter bubbles" in which conflicting viewpoints that do exist are largely filtered out of one's information diet thus fabricating the appearance of consensus. This fabrication further feeds into the human proclivity to subscribe to those beliefs that are also held by many others (Pariser, 2011; Miller & Record, 2013; Jiang et al., 2019; Anderson, E., 2021).

In our modern technological age, much of the information people receive from external sources is either technologically mediated (like information acquired via social media platforms or through targeted search engines), technologically produced (like content generated by bots), or both. Accordingly, working towards the preservation of epistemically secure societies will require attending closely to how people can and should justify their belief in information acquired from these sources. The epistemic security report culminates in a series of recommendations for pursuing epistemic

[4] One would be right to point out that human politicians also manipulate public trust quite successfully. My point is not that this is a unique capacity introduced by information producing and mediating AI technologies, but that AI technologies should not be expected to be limited in the same way as humans in this respect. While telling a convincing lie or presenting oneself as trustworthy takes skill and practice for humans, an AI system can be designed to optimize for gaining user trust.

security in technologically advanced societies, a key theme being that identifying and navigating the multitude of technologically enabled epistemic threats must be a multidisciplinary effort. Psychologists, sociologists, and political scientists, journalists and experts in media studies, historians, technologists, and philosophers of science and technology all have valuable perspectives to lend as to how people consume and evaluate information and justify their beliefs. This dissertation is a philosophical contribution to a new and important piece of the puzzle. I will discuss the impacts of one rapidly proliferating information producing technology – artificial intelligence (AI)-enabled expert systems – on epistemic security.

1. BELIEVING AI-ENABLED EXPERT SYSTEMS

This dissertation is about AI-enabled expert systems. AI-enabled expert systems are technologies designed to operate in roles traditionally filled by human experts, for instance by providing medical diagnoses and recommending treatment plans, sorting through job applicants, or approving loans. For example, pDOC is an AI-enabled expert system that processes clinical data and live brain scans to forecast the recovery of comatose patients (Song *et al*., 2018). The processes that underpin pDOC's predictions are opaque to humans such that human users are unable to trace in detail the processes by which the system arrives at its conclusions, yet the technology is slated to inform treatment planning and end-of-life decisions.[5] Given the high stakes of an incorrect prediction, the safe and responsible adoption of such a system will require that users have a way to evaluate when to take the system at its word, and when to second guess its outputs.

AI-enabled expert systems are slated to revolutionize a variety of fields ranging from medical diagnostics and treatment planning to financial services, legal proceedings, resource distribution, education and more. Accordingly, the central question guiding my thesis is as follows: As AI takes on roles traditionally filled by human experts, how can we – present and future AI users – acquire a degree of epistemic justification for our beliefs in the outputs (predictions, recommendations, diagnoses etc.) of the AI-enabled expert systems we are projected to use so heavily? In this dissertation I argue that we may do so in much the same way that a novice might derive justification for her belief in claims and advice of human experts, for instance by appraising explanations offered by the expert/AI or by looking for signs that the expert/AI is trustworthy.

My proposal that a person might appraise an AI system in the same way as she might appraise a human expert probably seems strange in light of the last couple pages; I just finished illustrating how the quick heuristics people use to evaluate human sources of testimony are ill-suited to appraising modern information producing and mediating technologies. The reason I compare AI systems to human experts stems from a particular similarity in the kind of relationship between human experts and those who seek their help and advice (novices) and between AI-enabled expert systems

[5] This definition of epistemic opacity is borrowed from Paul Humphreys (2004, 2009).

and users. Both are *epistemically imbalanced relationships* in which the information source (expert/AI) is epistemically privileged relative to the novice/user; she/it can utilize capacities, resources, and reasoning techniques to draw conclusions that the novice/user would be unable to access, reproduce, or in some cases, comprehend on her own.[6] For example, a physician's specialized knowledge and skill honed through years of training and experience places her in an epistemically privileged position relative to a lay patient who seeks the physician's advice. In the case of AI-enabled expert systems, the system's computational power allows it to identify patterns and draw conclusions from vast sets of data in a way in which no human is capable. I discuss what AI systems are and how they work in greater detail in section 2 of this chapter.

The interesting and problematic thing about epistemically imbalanced relationships is that when the epistemically disadvantaged party seeks out expert(-system) aid, then in virtue of the novice's epistemically disadvantaged position, she is not well-equipped to independently confirm the expert(-system's) response. Consider for example, a physician who outlines a cancer treatment regime to a patient. If the physician were then to try and explain to the patient how she decided upon that specific regime (including drug doses, timings, etc.) it is not clear how the explanation would help the patient justify his belief in the physician's claims. If an expert outlines her reasoning in such detail that it provides strong evidence in support of her claim – for instance, such that a series of true premises logically leads to a conclusion – then the novice is unlikely to have the expertise necessary to recognize the evidence as supporting the claim. We face a version of Meno's paradox: if a novice did hold such expertise, she would (a) not be a novice but an expert and (b) would have little need of expert advice in the first place.

Accordingly, the question stands, how can the novice, while remaining a novice, acquire justification for her belief in an expert claim? [7] A similar question can be asked of user-AI interactions: How can an AI user, without becoming an expert in the domain to which the AI system is applied, justify her belief in AI outputs? If an answer can be provided in the expert-novice case, then it would seem we would at least be on our way to acquiring an answer for the AI-user case. The issue

[6] I take the idea of relative epistemic privilege from Elizabeth Fricker (2006) who defines an expert as one who occupies an epistemically privileged position, whether it be through learning, skill, or luck, relative to another. There are, of course, other approaches to defining expertise. For instance, Elizabeth Anderson (2011) describes expertise as one's standing in a continuum of holders of domain-specific knowledge ranging from novices to field leaders; Alvin Goldman (2001) holds that expert standing requires one to meet a minimum threshold of knowledge and skill; and Hubert Dreyfus and Stuart Dreyfus (1986) argue that expertise necessitates a capacity for intuition – the ability to unconsciously extrapolate conclusions from experience and related domains of knowledge. For the purpose of this paper I am concerned with expertise only insofar as it describes the epistemically imbalanced relationships between naïve information recipients and epistemically well-placed information sources that are typical of novice-expert testimonial relationship. Therefore, I adopt Fricker's looser notion of expertise as relative epistemic privilege.

[7] Alvin Goldman (2001) poses a similarly phrased question with respect to how a novice might justify her decision to trust one putative expert over another: "Can novices, while remaining novices, make justified judgments about the relative credibility of rival experts? When and how is this possible?" (Goldman, 2001: 89). Numerous other scholars discuss epistemological challenges of expert-to-novice testimony. *Inter alia* see Fricker (2006), Anderson (2011), and Hardwig (1985).

is, however, that despite significant attention given to issues of expert-novice communication within literature on the epistemology of expert testimony and public communication of science, there is still much debate as to how, and to what extent, it is possible for novices to justify their beliefs in expert claims.[8] As such, turning to the more familiar question of justified belief in human experts does not provide quick answers to parallel questions about justified belief in AI outputs.

Accordingly, this dissertation has a dual purpose: The primary objective is to investigate how AI users can acquire a degree of justification for their belief in AI outputs by using the novice-expert model as a guide to help identify and frame key challenges to user appraisal of AI systems. However, this primary objective is achieved while pursuing a secondary goal of addressing standing questions about the justification of novice belief in human expert claims along the way.

The remainder of this introductory chapter proceeds in five sections. The next two sections briefly attend to terminology. In section 2, I describe what Artificial Intelligence is, and I clarify which kinds of AI systems are of primary concern for this dissertation. In section 3, I turn to justification. I explain what I mean by degrees of justification and comment on the relationship between internal and external justification. In section 4, I expand on the novice-expert model for the user-AI relationship, and I defend the model from complaints that likening AI systems to human experts inappropriately anthropomorphizes AI technologies. In section 5, I discuss the appraisal of AI/expert past performance as a method for justifying one's belief in AI outputs and expert claims, and I argue that even where information about expert/AI track record is readily available, it may still be important to pursue other reasons to underpin belief in a specific claim or output. Section 5 then segues to section 6 in which I introduce subsequent chapters.

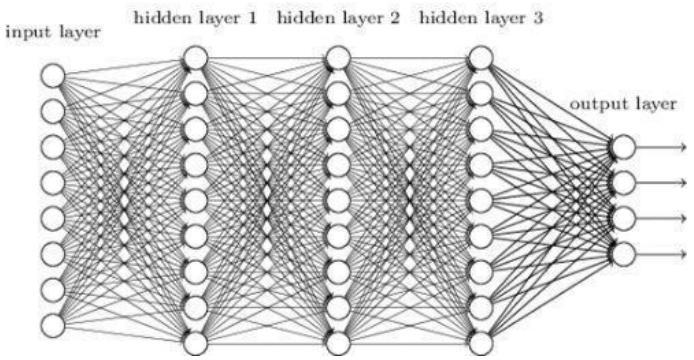
2. WHAT IS ARTIFICIAL INTELLIGENCE?

Artificial Intelligence (AI) generally refers to computational systems capable of performing tasks that typically require human intelligence such as driving cars, diagnosing medical conditions, giving investment advice, and so forth. AI describes a wide range of computational technologies used to mimic human behavior and judgment (Russel & Norvig, 2010; Franklin, 2014). The simplest AI systems, called symbolic AI or Good Old-Fashion AI (GOFAI), utilize logic- and knowledge-based approaches to encode human knowledge by, for example, building decision trees (Boden, 2014). On the other hand, more recent advances in AI technology utilize machine learning approaches to teach systems how to autonomously classify data or generate original outputs without using a complex set of explicit rules. Machine learning approaches include supervised learning techniques in which an AI system is trained by exposing it to large quantities of prelabeled data, reinforcement learning

[8]Alvin Goldman (2001), Elizabeth Anderson (2011), Walton (1989), Reiss (2008), Hume (1748) all outline taxonomies for novice evaluation of expertise. Carlo Martini (2014) provides a thorough overview. Axel Gelfert (2014, Chapter 9) provides an informative introduction to the literature on expert testimony.

techniques in which a system learns by being 'rewarded' for correct outputs and 'punished' for incorrect outputs, and unsupervised learning techniques in which a system is fed large quantities of unlabeled data in which it finds patterns using brute statistical processing (James et al., 2013).

The most complex machine learning methods utilize deep learning. Deep learning (or hierarchical learning) is a machine learning method based on the construction of multilayer artificial neural networks (ANNs) inspired by the biological neural networks of the animal brain (Goodfellow et al., 2016). ANNs are comprised of artificial neurons, or nodes, and the connections between the nodes called edges [figure 1]. Nodes receive and process input signals from previous nodes and produce outputs. Each edge is assigned a weight that represents the relative strength of the connection between specific nodes, similar to how the connections between some neurons in the animal brain are stronger than others. The greater the weight, the greater the influence one node has on the following node. The first layer of nodes, called the input layer, is designed to receive external data such as images or documents. The nodes in the hidden layers receive and sum up the weighted outputs of the previous layer (weighted according to the weight assigned to the edge) and only fire (send new outputs to the following layer) if a certain threshold value is met. The output layer produces the ultimate outputs in the form categorized data, identified images, etc. The ANN learns to be a better classifier of data by incrementally adjusting the various node thresholds and edge weights to minimize the observed error rate (e.g. the rate of incorrect image classifications).



[Figure 1] Deep Artificial Neural Network.
Nodes and edges are depicted as circles and arrows respectively.

In this dissertation I limit my discussion on AI in a few ways. First, I focus my discussion to AI-enabled expert systems – AI systems that operate in roles traditionally filled by human experts – that produce outputs which people use to inform further decision-making and action. Examples include medical diagnostic or treatment systems like pDOC, financial RoboAdvisors, and recidivism prediction software. I do not focus on AI systems that surpass the human decision point to also *act* autonomously like self-driving cars, autonomous weapons, or financial trading software. I assume that if a person were to reject a system's output because she thinks it is false or believes the system to be

unreliable or untrustworthy, then it is also safe to assume that for an analogous system that skips the human decision point, the user would also reject the services of the autonomous system given that she would not have deferred to the system's outputs had she been looped into the decision point. If, for example, an autonomous trading software trades stock at a price that the human user would not agree to trade her stocks at if she were looped into the decision, then we may also assume that the user would not trust the system to trade her stocks autonomously.

Second, I am primarily concerned with AI-enabled expert systems that employ the more complex machine learning techniques described above. These techniques underpin the most recent endeavors in AI research and development and are more powerful and able to mimic more complex and cognitively demanding behaviors than symbolic AI.

Furthermore, AI systems with complex internal processes are epistemically opaque, meaning that humans are unable to trace in detail the processes that lead from the input of the device to the output (Humphreys, 2004, 2009). As Mark Coeckelbergh (2019) explains:

> In the case of these machine learning applications, it is not clear how exactly the AI system arrives at its decision or recommendation. It is a statistical process and those who create it know in general how it works, but even the developers — let alone further users and those affected by the algorithm — do not know how the system arrives at a particular decision relevant to a particular person. They cannot explain or make transparent the decision making in all its steps.

Much of the discomfort around widespread adoption of AI technologies stems precisely from the concern that there is no designer, developer, or researcher who can trace in detail how a machine learning system derives its outputs and, in turn, use that knowledge to determine if and how the system is prone to error.

More so, key to this dissertation is the observation that an AI system's epistemic opacity establishes a strong parallel between expert testimony and AI outputs as potential knowledge sources. Human experts are also often epistemically opaque to novices. For instance, an expert's expertise might derive from her well-honed intuitions (Dreyfus & Dreyfus, 1986), in which case the expert would be unaware of her own reasoning process let alone able to communicate it to a novice. Alternatively, where an expert does consciously reason to her conclusions, in virtue of the epistemically imbalanced relationship between expert and novice, the novice would not hold the background knowledge and/or cognitive capacities to comprehend the expert's full reasoning process so as to determine whether the expert's claim is epistemically responsible. In this way, as Paul Humphreys (2009) notes, "there are parallels between testimonial evidence from other people and our reliance on outputs from computational devices. In both cases we do not have direct access to the source of the evidence but must rely on the authority of an intermediary" (227). Humphreys does not expand on what he means by "the authority of the intermediary," but I presume he refers to deciding whether the AI/expert is a kind of information source that ought to be believed. Accordingly, the

challenge in both cases is determining what makes an AI/expert the kind of thing that ought to be trusted. I address the topic of trust in AI and experts in chapters 3, 4, and 5.

One would be correct to note, however, that it seems unnecessary for a user or novice to have to appraise the entirety of an AI system's output derivation process or an expert's reasoning processes to acquire some sense of the veracity or epistemic well-foundedness of the AI's output or expert's claim. A human physician, for instance, can provide a patient with a lay-accessible explanation for a diagnosis. Similarly, research into explainable AI (XAI) aims to develop strategies for developing AI systems that can provide users with understandable explanations for AI outputs (Adadi & Berrada, 2018; Gilpin et al., 2019). For example, a loan recommendation system might explain its decision to reject a loan request with a counterfactual explanation listing the smallest changes a subject would have had to make to her profile in order to receive a different outcome; the system might identify a specific increase in income or property holdings or a change in spending habits that would have led the system to recommend approval. Such an explanation does not provide a direct window to the inner workings of the AI system like one might acquire by lifting the hood of a car to observe the running engine, but it does provide a post hoc account of the system's 'reasoning' that is intelligible to users. However, that an explanation may depart from the original processes employed to derive an output raises an interesting question: How does an explanation's abstractions from an original derivation process impact a recipient's ability to use the explanation to help justify her belief in an output? In other words, does an explanation help a recipient to justify her beliefs because of or despite its abstractions? Chapters 1 and 2 respond.

3. DEGREES OF INTERNAL JUSTIFICATION

In this dissertation I investigate how novices and AI users might acquire a degree of what I call "internal epistemic justification" for their beliefs in AI outputs and expert claims. In this section I first clarify my emphasis on "degrees" of epistemic justification and then defend my internalist perspective.

Degrees of Justification

Knowledge is often understood in the traditional sense as a justified true belief. When understood as such, knowledge is obtained when a belief is both true and an acceptable threshold of justification for the belief is met. Accordingly, epistemology – the study of knowledge – is largely geared toward investigating what threshold of justification is needed for a belief to in fact 'be justified' and how that threshold can be satisfied.

However, while knowledge is binary – either a belief is both justified and true or it is not – justification is a matter of degree. In this dissertation I am not concerned with knowledge acquisition and I therefore do not investigate what threshold of justification is needed for a belief to be

sufficiently well justified to constitute knowledge. Rather, I adopt a graded view of justification that aligns with the intuitive notion that people can have a range of poor to very good reasons for believing what they do. The idea is that the better the reasons (justifications) they hold, the more likely their beliefs are to be true.

There is a challenge, however: people sometimes hold, or think they hold, good reasons for their beliefs when those beliefs are in fact false. For example, in the morning I find the street outside my house is wet, and I form the belief that it rained overnight. This is a reasonable conclusion given that the entire street appears to be covered in water; the same appearance would not be produced by a lawn sprinkler, a morning frost, or even a burst water main. What really happened, however, was that the city street cleaner drove past just before dawn and pressure washed the tarmac. The potential for a mismatch between a seemingly well-justified belief and the veracity of such a belief, as in the example above, fuels a debate between internalist and externalist theories of knowledge and justified belief.

Internalism v. Externalism

Internalism and externalism are perspectives regarding what conditions and events can contribute to epistemic justification for belief.[9] Internalists hold that a person must have internal access to the justifiers for a belief for that belief to be justified (Steup, 1996; Conee & Feldman, 2004). As Matthias Steup (1996) writes:

> What makes an account of justification internalist is that it imposes a certain condition on those factors that determine whether a belief is justified. Such factors – let's call them "J-factors" – can be beliefs, experiences, or epistemic standards. The condition in question requires J-factors to be internal to the subject's mind or, to put it differently, accessible on reflection.

Externalists, on the other hand, hold that the factors that make a belief justified exist external to the believer such that they are beyond her awareness (Goldman, 1999, 2008; Plantinga, 1993). Alvin Goldman is perhaps the most well-known defender of externalism. He is known for proposing a form of externalism called *external process reliablism* whereby an agent is justified in a belief if the belief is resultant from a reliable belief formation process. It matters not whether a believer holds any reason to think her belief forming process is likely to yield true beliefs. For example, if a person forms a belief about the world via visual perception – e.g. I believe there is a barn in the field because I see a barn in the field – and perception is a reliable process by which to form beliefs, then a belief formed via that process constitutes knowledge – e.g. I know there is a barn in the field. Goldman (1976) notes, however, that a situation may arise in which perception is not a reliable process by which to form beliefs. For, example if taking a ride through a stretch of countryside where many structures

[9] See BonJour (1992), Audi (1998), and Plantinga (1993) for extended analyses of the debate between internalist and externalist conceptions of justification.

which appear to be barns are in fact bard façades, one's belief formation process about the presence of barns via perception would not be reliable. Therefore, in this stretch of countryside one's belief about there being barns in the field would not constitute knowledge regardless of how internally justified one feels in her belief. In this way, externalist theories like Goldman's provide a way to clearly delineate between beliefs that constitute knowledge and beliefs that do not.

I am not, however, concerned with knowledge at this time, and I therefore remain agnostic as to whether a person can be epistemically justified in a belief without some internal awareness of her reasons. I am presently concerned with practical issues of informed decision-making, specifically how one informs her decisions about whether to believe the claims and outputs of experts and AI systems. If a person engages in informed decision-making it implies that she holds internal reasons for her decision (to believe), and that her decision (to believe) is not made on a whim. Externalist theories, however, are, by definition, not conducive to discussion regarding how a person internally informs her decisions to believe. Therefore, I adopt an internalist conception of justification according to which an individual's degree of justification turns on the reasons she holds and the inferences she makes.

Externalists, however, are quick to point out a key issue with internalist theories of justification; it is an issue that plagues internalism whether discussing justified belief for knowledge acquisition or seeking out degrees of justification for informed decision-making. The challenge is that decisions can be terribly *mis*informed, and therefore it is possible for a person to feel internally justified in holding beliefs that are in fact false (Feldman, 2003; Lemos, 2012). I address this grievance in greater detail in Chapter 2, but for the time being I would like to emphasize that this issue is not a reason to dismiss the internalist perspective I adopt. Rather, the criticism helps illustrate the severity of the challenge this dissertation seeks to address. Ideally, a novice or AI user would hold a degree of internal epistemic justification for belief in expert/AI outputs and recommendations which are also externally justified. Figuring out how to approach this ideal is the tricky part. Novices and AI users are in epistemically disadvantaged positions relative to their expert and AI informants which means they lack, to varying degrees, the relevant background knowledge and/or cognitive capacities needed to ensure such decisions are well-informed, not misinformed.

4. UNNECESSARY ANTHROPOMORPHIZATION

Having proposed that user-AI interactions be investigated as analogous to epistemically imbalanced expert-novice relationships, I will now preempt the complaint that likening AI systems to human experts unnecessarily anthropomorphizes the technology. As advances in AI research extend the capacities of AI technologies into areas once considered the preserve of human intelligence, it is increasingly common for anthropocentric psychological terms such as awareness, perception and autonomy to be used to describe system behavior (Shevlin & Halina, 2019). David Watson (2019: 417) points out that even "the name of the discipline itself – artificial intelligence – practically dares

us to compare our human modes of reasoning with the behavior of algorithms." However, there is concern that close analogy between human and machine intelligence may lead to misunderstanding of AI capacities (Shelvin & Halina, 2019; Wortham, Theodorou & Bryson, 2016). For instance, Watson (2019: 417) asserts that "such rhetoric is at best misleading and at worst downright dangerous."

There are several reasons people caution against the anthropomorphization of AI technologies. One reason is the exaggeration of hopes and fears that surround the technology and its uses. Owing to the use of physiological terms to describe AI behavior and the sensational (and often humanoid) representation of AI in popular media – news and science fiction – the terms "AI" and "artificial intelligence" often evoke images of both dystopic and utopic scenarios (Cave, Coughlan & Dihal, 2019; Cave & Dihal, 2019). Exaggerated fears include "Terminator"-like scenarios in which self-aware AI systems make selfish decisions to value their own existence and autonomy over human life, while exaggerated hopes entertain the possibility of using AI to extend human life (or the experience of life) past normal life expectancy or indefinitely (Cave, Coughlan & Dihal, 2019; Bryson & Kime, 2011). David Watson (2019) and Joanna Bryson and Philip Kime (2011) argue that the problem with prevalent sensationalism about human or more-than-human AI capacities is that it prevents us from properly conceptualizing the ethical challenges posed by AI and distracts from real and more immediate dangers. For instance, Bryson and Kime contend that the most pressing dangers posed by AI are not due to any unusual or fantastic AI capacities, but by the people who use AI systems; AI systems are like any other artifact – like guns, books, or governments – in that they can be used to benefit humans, or they can be misused to our detriment, either unwittingly or with malicious intent. So to summarize, the fear is that the anthropomorphization and associated sensationalization of AI technologies steers us away from discussing the human factor of how AI systems are used and to what purposes they are employed.

A second reason to avoid unnecessarily anthropomorphizing AI is to prevent instances of unwitting AI misuse. Even if a person does not entertain the more exaggerated AI hopes and fears, false assumptions about human-like cognitive capacities can lead to improper employment of an AI technology. For example, compared to AI systems, humans are cognitive generalists blessed with what Nicholas Carr (2015: 121) describes as "our [human] ability to make sense of things, to weave the knowledge we draw from observation and experience, from living, into a rich and fluid understanding of the world that we can then apply to any task or challenge." This ability allows humans to use their existing knowledge to draw conclusions and react to scenarios they have never before encountered. For example, a human physician would know from training and experience that asthma is a respiratory condition that causes lung tissue inflammation, constriction of bronchioles, and difficulty breathing. Accordingly, even if the physician had never encountered a pneumonia patient that suffered from asthma before, it is reasonable to assume that she would be able to extrapolate from her knowledge about asthma that a pneumonia patient with asthma would be at higher risk of complications and death than a patient without.

However, unlike humans, AI systems operate in narrow domains of expertise strictly defined by the data sets on which they are trained. There is consequently a risk that when AI-enable expert systems are employed to perform tasks traditionally performed by human experts, that anthropocentric terms like *intelligence* and *expert* prompt users to unintentionally project qualities of human cognitive generality onto AI systems which, in turn, can result in human failure to recognize narrow contexts in which an AI system can operate. For example, Alex London (2019) describes Rich Caruana et al.'s (2015) account of a medical AI system that learned to predict lower risks of death due to pneumonia for asthmatic patients than for the general population. The system was trained on hospital admission and patient outcome data. However, because human medical experts know that asthmatic pneumonia patients are at risk for fatal complications, these patients have historically been admitted directly to the Intensive Care Unit (ICU) for observation. The additional care received in the ICU in turn lowered the relative pneumonia-based death rate of asthmatic patients which led the expert system to predict, based on the historical data, lower risk of death for asthma patients. London points out that the expert system did not execute its function poorly. To the contrary, the system performed perfectly according to its training. However, the human users and designers failed to properly identify the system's precise domain of expertise as defined by the patient data on which it was trained and tested – predicting the mortality of pneumonia patients *given* the patients have received current standards of care. They had instead unknowingly assumed that like a human physician endowed with some degree of cognitive generality, that the expert system would predict the mortality risk of pneumonia patients, full stop.

Overall, the proper use of AI technologies requires that users are able to correctly identify AI system capacities and respond accordingly. However, anthropomorphization of AI can mislead human intuitions about potential and realized AI capacities which, in turn, can result in unintentional system misuse and distract from real and pressing risks that the adoption of AI technologies does pose. It is therefore good practice to avoid using anthropomorphic and psychological terminology in describing AI in order to prevent AI sensationalism and the assumption of false equivalencies between AI systems and human minds. AI systems should be thought of only as the mere artifacts – in this case computational instruments – that they are (Bryson 2010, 2018a; Bryson & Kime, 2011).

I largely agree that over-anthropomorphization of AI should be avoided. However, I would also caution against swinging too far in the opposite direction. The flip side to resisting comparison between humans and AI is ignoring instances where familiar problems in the human case have simply taken on a new form, or better yet, where those familiar problems may benefit from a new perspective when framed as a challenge to AI as well. Therefore, I propose that invoking analogies between AI systems and humans is appropriate to the extent that the analogy is relevant to illustrating and addressing a common challenge. Accordingly, in this dissertation it is not my aim to anthropomorphize AI by comparing AI-enabled expert systems to human experts. Rather my primary focus is on the similar challenges posed by the epistemically imbalanced nature of both user-AI and human novice-expert relationships. If I can make headway into understanding how novices might

acquire a degree of epistemic justification for their belief in expert testimony, then it is possible similar conclusions might be drawn about how users can acquire a degree of epistemic justification for their belief in AI system outputs. Conversely, lessons learned by looking into the AI case may help address standing questions posed by the human expert-novice case.

5. TRACK RECORD

Finally, I will address the complaint that emphasizing the challenges of epistemically imbalanced relationships between users and AI systems, and between novices and experts, introduces an unnecessary complication to the issue at hand. There is an argument that the best way to inform one's decision about whether to believe an instrument's outputs (e.g. readings or predictions) is to review its track record. Track record provides evidence of reliability, where reliability is understood as the characteristic of something or someone that consistently performs in the way one would expect or hope.[10] If track record is available, then there is no need to look for further reason to rely. For example, Michael Bishop and J. D. Trout (2002, 2005) defend the idea that track record *alone* justifies deferring to the advice of statistical prediction rules (SPRs) – weighted statistical models used to predict outcomes – over human expert advice. SPR-enabled expert systems have been shown to consistently outperform human experts in predicting student performance, employee retention, violent crime, medical prognosis, and so forth, though it is difficult to explain why SPRs enjoy their success. Bishop and Trout (2005: 16) write:

> Even if there is no good explanation for their relative success, we ought to favor [SPRs] over human judgment on the basis of performance alone. After all, the psychological processes we use to make complex social judgments are just as mysterious as SPRs, if not more so… It might be that given our current understanding, replacing human judgment with an SPR may inevitably involve replacing one mystery for another—but the SPR is a mystery with a better track record.

I agree that as evidence of reliability, track record can play a very important role in providing justification for belief in AI systems and human experts alike. Indeed, it may even be the case that the more information that is available about past performance, the less important other sources of justification for belief – such as evaluations of reasoning processes or explanations thereof – become. But if it is the case that track record alone can always be used to determine whether an expert or AI is a reliable source of information and therefore ought to be believed, then this dissertation might as well end here: a person acquires epistemic justification for belief in both human expert claims and AI outputs by looking to track record for evidence of expert/AI (un)reliability. It does not matter, for instance, that an AI system is epistemically opaque to users or that a novice is not equipped with the expert knowledge and/or cognitive capacity necessary to make meaningful appraisals of expert

[10] See Chapter 4 for an extended discussion on the definition of reliance and the relationship between reliance and trust.

reasoning processes or explanations thereof (see chapters 1 and 2). I argue, however, that there is still good reason to investigate these sources, as I do in this dissertation.

First, there is the issue of track record availability. For newly developed instruments or newly trained human experts, track record may not yet be well-established. The fewer test cases that underpin a reported track record of success, the less weight that should be assigned to the track record in evaluating reliability in any one instance, and the more important other sources of justification become.

Second, even where track record is well-established it cannot be assumed that past performance is a conclusive test for future performance. Therefore, additional factors that can help predict future performance should also be taken into account. Daniel Hausman (1992) illustrates this point with the example of a used car. If a mechanic were to look under the hood of a car and note the condition of its internal components, then the mechanic would hold information relevant to the car's future performance including if and when the car might experience a mechanical failure and how it is likely to perform under different conditions (e.g. hot weather, steep up-hills, or driving through snow). Regardless of how many years the car has functioned without incident, it would be absurd for a car owner not to take the mechanic's observations into account when deciding whether to rely on the car for future drives. Indeed, if track record were an infallible indication of future performance, then there would have been no reason to look under the hood in the first place. Accordingly, Hausman holds that while it may be the case that "the more information we have about performance, the less important is separate examination of components... it remains sensible to assess assumptions or components, particularly in circumstances of breakdown and when considering a new use" (72).[11]

Of course, human intellect and AI algorithms do not wear out in the same way as mechanical tools. However, the point that track record cannot always be assumed to extrapolate to future cases still holds, especially with regard to the extrapolation of track record to predict performance in new or dynamic contexts. For example, a human physician newly employed to a hospital that uses cutting edge medical instrumentation may not perform as well in her new place of work if she is not experienced with the tools she is expected to use. The same is true of AI systems. For example, IBM's Watson for Oncology – an AI-enable expert system built to aid in treatment planning for cancer patients – was primarily trained on data sourced from Memorial Sloan Kettering Cancer Center where the typical oncology patient is white and middle-class and where medical practices are cutting-edge. When rolled out to hospitals in India and Southeast Asia that observe different medical practices and

[11] Donald Gillies (2016) makes a similar argument in favor of using both statistical evidence (evidence collected from randomized control trials (RCTs)) and evidence of mechanism in evaluating the safety and efficacy of new medical therapies. The evidence-based medicine (EBM) movement holds RCTs as the gold standard for safety and efficacy evaluations, however Gillies present two case studies – one regarding the use of streptomycin as a treatment for tuberculosis and a second regarding the use of thalidomide to treat morning sickness – in which concurrent consideration of both statistical evidence and evidence of mechanism did (or would have, in the case of thalidomide) prevented erroneous conclusions to be drawn about therapy safety and efficacy.

serve different demographics, Watson performed unsatisfactorily due to its different training context (Ross & Swetlitz, 2017). The Watson for Oncology example is rather obvious. It probably should have been predicted that Watson would perform differently with such a significant change in environment. However, as illustrated by Alex London's (2018) discussion of the medical AI system that predicted low risk of mortality for pneumonia patients with asthma, the context in which an AI system operates can be quite narrow, so narrow in fact, that both an AI system's users and designers can struggle to identify the bounds within which the system should be expected to operate-well (also see Martini, 2014). Accordingly, as Hausman (1992: 72) writes, "the fact that a computer program [(in our case, AI system)] works in a few instances does not render study of its algorithm and code superfluous or irrelevant." It is still important to look under the hood, to crack open the black box, and take a peek at what is going on inside.

Finally, a third reason to look beyond past performance is that even where track record is well-established and issues of degradation and context sensitivity are set aside, it is often the case that track record is not clearly communicated to those looking to justify their decisions to believe. For example, when visiting a new physician, a patient is not usually provided with a report of the physician's past successes and failures. Rather, an acceptable consistency of good past performance is alluded to by certifications, awards, and the physician's maintenance of a medical license. Similarly, the reliability of instruments including AI-enabled expert systems might be indicated by testing certificates. Such proximate indications of track record are important as lay patients and users often will not have the necessary expertise to interpret more detailed performance reports themselves. However, the belief that certifications and awards do indeed stand as good proxies for track record also requires justification. What reason does a person have to believe that those people or groups who issue certification are themselves reliable identifiers of reliable experts and instruments? In turn, who certifies the certifiers?[12] Track record proxies only serve as meaningful evidence of reliability if backed by well-functioning systems of appraisal and quality control. Accordingly, the reason most people have for believing experts, AI systems, and other instruments does not reduce to track record but to what I describe in Chapter 5 as systemic trust – the kind of trust a person holds in an expert or expert system because she/it is embedded in a larger social-epistemic system that regulates trustee performance.

I reiterate that it is not my aim to say that track record is unimportant. To the contrary, where it is available, track record should be taken into consideration when deciding whether to (dis)believe the outputs of AI systems and human experts. However, given possible issues with the extrapolation of past performance to current scenarios, and given the indirect ways in which people typically

[12] Alvin Goldman (2001) refers to those who are experts at appraising other experts as meta-experts. Here I point out the need to consider not only meta-expert appraisal, but also meta-meta-expert appraisal, and meta-meta-meta-expert appraisal, and…

acquire information about the track record, it is also important to consider other ways in which a person might buttress her justification for believing expert or AI claims.

6. WHAT FOLLOWS

This dissertation proceeds in two parts. Part 1 on AI and expert explanations investigates how explanations allow a person to 'look under the hood' to see how experts and AI systems derive their outputs. The overarching goal of Part 1 is to investigate what I call the *justificatory value* of explanations, where an explanation is justificatorally valuable insofar as it conveys information that, via coherence with an explanation recipient's background knowledge and beliefs, helps the recipient to acquire some degree of internal epistemic justification for her (dis)belief in the explainer's claims or outputs.

Within Part I, Chapter 1 first clarifies a philosophical approach to discussing AI explanations. For philosophers looking to better understand what makes for a good AI explanation, philosophy of explanation may seem an obvious place to start. After all, "explanation" is in the title of the philosophical discipline. I caution, however, that the paradigm structure of an explanation as understood by philosophers of explanation can be an awkward model for thinking about the justificatory value of explanations offered by AI systems. Instead, I argue that AI explanations are better understood like the kind of post hoc and partial explanations offered by experts to novices for their reasoning processes. These explanations do not provide direct windows to the inner workings of an AI system like one might acquire by lifting the hood of a car, but they provide a post hoc account thereof, like a report provided to a car owner by a mechanic or an explanation for a diagnosis provided by a physician to a lay patient. As such, there is a question as to whether the explanation is justificatorally valuable because of or despite its abstractions from the original reasoning process. In the second half of this chapter, I discuss how an explanation's simplification or fidelity to an original reasoning process influences the justificatory value of an explanation.

Chapter 2 then tackles the root of the problem: are expert-to-novice and AI-to-user explanations justificatorally valuable? If so, how much? Considering the cognitive capacities and relatively limited background knowledge of epistemically disadvantaged novices and lay users, my conclusion is largely a negative one: the more epistemically imbalanced a relationship, the less justificatorally valuable an explanation can be.

Accordingly, Part II changes gears. Where track record of expert/AI performance is unavailable or insufficient and explanations are of low justificatory value, a novice might instead base her decision to believe an expert claim or AI output on epistemic trust. Instead of attempting to directly evaluate the quality of an expert claim or AI output, one may indirectly acquire a degree of

epistemic justification for her beliefs by appraising whether the information source is one that is likely to produce true or epistemically well-founded outputs.

Chapter 3 explores the nature of epistemic trust in experts and in AI, arguing that both human novice-expert trust relationships and user trust in AI are best understood in terms of what I call "systemic" as opposed to "idiosyncratic" trust. While idiosyncratic trust is epistemic trust held in an individual because of what one believes about an expert's individual characteristics and features, systemic trust is epistemic trust in an expert grounded in a novice's belief that an expert is a member and/or product of an expert community and larger social-epistemic system that influences the expert's behavior and performance. I caution that discussions that frame challenges to novice/user trust in experts/AI as issues of idiosyncratic trust (as they most often do) misdirect practical conversations about evaluating trustworthiness and attracting user/novice trust.

Chapter 4 briefly interjects to address concerns about applying trust terminology to artifacts (like AI systems or other computational and mechanical tools or instruments) and to present my own theory on the differentiation between the terms "reliance" and "trust" and a third term, "use".

Chapter 5 concludes Part II by expounding on the foundations of systemic trust in both human experts and in AI-enabled expert systems by investigating what features make for epistemically well-functioning systems. I understand an epistemically well-functioning system as one that both attracts novice trust to embedded experts and renders that epistemic trust in experts well-placed, and I describe two premises – the sociological premise and the enhanced epistemological premise – which, when satisfied, describe an ideally well-functioning system. I evaluate the state of systemic trust in AI according to these two premises. The chapter concludes with recommendations for improvement.

Overall, my aim in this dissertation is to investigate how the deployment of AI-enabled expert systems across various areas of society impacts epistemic security – that is, how these systems impact the ability of people to differentiate true information and epistemically trustworthy information sources from false information and epistemically untrustworthy information sources. A main theme of the discussions that follow is that challenges to epistemic security posed by AI-enabled expert systems are not much different than those posed by human experts, but that the challenges posed by human experts are not necessarily easy challenges to address.

PART I

EXPLANATION

1

AI and Expert Explanations

AI enabled-expert systems are slated to inform high-stakes decisions such as diagnosing medical conditions, approving loans, granting parole, hiring employees, organizing the distribution of scarce resources, and predicting crime. With the increasingly consequential roles AI systems are being proposed to perform, worries about the safe and responsible use of AI technologies grow, and it is becoming more important that users are able to determine when to believe AI outputs and recommendations and when it would be best to disbelieve.

Setting aside the option of looking to system track record (See section 5 of the introduction), a straightforward solution would be to appraise the system's reliability like one would appraise any other tool or instrument; lift up the hood, take a look inside, and see if the system is working as it should be. However, as prefaced in the introductory chapter, concerns about AI arise precisely from the fact that lifting the hood is not a plausible solution. While an AI system could easily be programmed to output reports of the exact algorithmic processes leading to each of its outputs, such a 'direct look' would be of little use to a human user. It is the central challenge of XAI literature that AI systems are epistemically opaque such that no person can trace in detail the processes that lead from inputs to outputs. For example, John Zerilli et al. (2018) describes deep learning, a common AI design technique, as "involving multiple hidden layers of processes that are fiendishly intricate and virtually impossible to unsnarl" (See also Burrell, 2016). If we were simply to open the black box – to look under the hood – we humans would not be able to comprehend what we see and therefore have no way of using that information to meaningfully inform our belief in the system's outputs.

The field of explainable AI (XAI) research has emerged to address challenges posed by AI system opacity. The general idea is that if algorithmic mechanisms of AI systems are incomprehensibly complex, then perhaps AI systems can be designed to produce cognitively accessible explanations for their outputs instead. Among other things, AI explainability strategies are expected to address issues of AI safety and control, to correct biases and ensure accountability in AI decision-making, and – my present focus – to help users justify their belief in AI outputs and recommendations by in some way elucidating how and/or why the AI system produced the output it did (Adadi & Berrada 2018; Weller, 2017).

The overarching goal of Part 1 of this dissertation is to investigate what I call the *justificatory value* of AI explanations. I consider an explanation to be justificatorally valuable insofar as it conveys information that, via coherence with an explanation recipient's background knowledge and beliefs, helps the recipient to acquire some degree of internal epistemic justification for her (dis)belief in the explainer's claims, conclusions, or recommendations. Briefly put, epistemic justification via coherence depends on how a belief supports, is supported by, or otherwise dovetails with existing beliefs (assuming, of course, that those beliefs are themselves true or epistemically well-founded. I discuss coherence theories of justification in more depth and grapple with this complication in Chapter 2) (Quine & Ullian, 1970; Feldman, 2003; Thagard, 2005; Lemos, 2012). Consider, for example, an explanation offered by a medical diagnostic AI system to a patient for the AI system's conclusion that a mole is not cancerous and does not require a biopsy to confirm. The AI offers the following explanation for its conclusion: "though the mole is raised and has a rough texture which can be an indicator of cancerous or precancerous skin blemish, it is evenly colored and has well-defined edges which are strong counter indicators of cancer." Assuming the patient has some minimal background knowledge about the visual representation of cancerous skin blemishes (for instance, that color and shape often factor into diagnoses), then the AI system's explanation might tell the patient that the system factored relevant information into its diagnostic process. In turn, the patient may have more reason to believe that the AI system's diagnosis is likely to be true than if the system's explanation seemed to reference irrelevant factors like the color of a person's wristwatch (also in the picture). There are certainly arguments to be made against this analysis, and I will address them in due course as it is the goal of this and the following chapter to determine just how justificatorally valuable we can expect such explanations to be.

Part I proceeds in two segments. This Chapter 1 first establishes a philosophical approach for investigating the justificatory value of AI explanations. I begin in section 1 by presenting a way of thinking about AI explainability – which I call "adopting the testimonial stance" – that likens AI explanations to those offered by human experts; AI and human expert explanations share similarities that pave the way for a mutually beneficial investigation of their justificatory value. Starting in on that investigation, in section 2, I appraise how the literature in philosophy of explanation might inform the discussion. This is an obvious place to start. Afterall, "explanation" is in the title of the philosophical

discipline. I caution, however, that the literature is multifaceted and can easily mislead if one dives in unguided. In section 3, I show how a particular discussion taking place in the literature on contexts of discovery and justification in scientific research provides just the framework we need to guide an investigation into justificatory value of both AI and human expert explanation. The discovery/justification literature is not often connected to issues of explanation, but I demonstrate that it provides a natural backdrop for the discussions that follow. Using the discovery/justification framework, in section 4, I identify and describe how two different factors – explanation detail and explanation reflectivity – can influence the justificatory value of an explanation.

Chapter 2 tackles the meat of the issue; I make my case for just how justificatorally valuable we can expect AI and human expert explanations to be.

1. THE TESTIMONIAL STANCE

Whatever the goal of XAI, there are two possible approaches to tackling AI opacity. The first option is simply to build less complex AI systems. The hope is that less complex AI architectures can more easily be made transparent and therefore more easily be appraised for sources of error.[1] This approach is plagued, however, by the so-called AI explainability-accuracy trade-off (Kuhn & Johnson, 2013; London, 2019; Duval, 2019; Gilpin et al., 2019); when AI explainability is understood in terms of system transparency, more explainable (more transparent / less opaque) systems will be less complex, however reductions in system complexity also reduce system's computational power. As such, less powerful systems are less accurate in their predictions, classifications, etc., and, inversely, any significant gains in AI system power (and therefore accuracy) will require gains in complexity. So, unless we are willing to ban the production of epistemically opaque AI systems and lose out on the potential benefits of more powerful computational tools, we must think about AI explainability in a way other than as striving for mere system transparency.

The second and more popular vein of XAI research therefore aims to develop strategies for providing users with understandable explanations for why AI systems produce certain outputs without reducing system complexity (Guidotti, 2018). The primary goal is not to provide a direct view to the original algorithmic processes by which an output was derived, but to provide an explanation recipient with the information she needs to achieve some other end, such as informing her decision to believe the output, showing her what changes would result in a different output (e.g. a higher credit score or a loan approval), or helping her appraise the system for sources of error or bias (Tomsett et al., 2018). Such explanations do not necessarily shed light on the original computational processes behind the output. For example, case-based explanations, also called example-based explanations, help a user

[1] AI system complexity can refer to a variety of system features. Some examples include the number of nodes, edges, and layers in a neural network (See figure 1 in the introductory chapter), the number of non-zero weights in the model, the depth of decision trees where decisions trees are used, the number of defined boundaries in the model, or the length of any rule conditions used in the model (Guidotti, 2018: 10).

understand why an AI system produced an output by showing the user particular instances in a data set to help account for the AI system's behavior (Bien, J. and Tibshirani, 2011; Kim et al., 2014; Gurumoorthy et al., 2017). For instance, a medical treatment planning AI system might point to a handful of medical journal papers that influenced its recommendations. Those particular articles are merely "prototypes" – representative examples – of the thousands of articles the system processed in deriving its conclusion. In this way, the example-based explanation misrepresents the process by which the AI system derived its treatment plan as a process of reasoning from example instead of statistical inference. Nonetheless, the information provided by the explanation might still be used by a physician to evaluate whether the AI system has processed the right kind of information to inform its recommendation, for example, whether the papers referenced by the system do indeed pertain to the patient's particular diagnosis or whether those papers consider how any of the patient's pre-existing conditions might affect treatment. I expand on this point further in Chapter 2.

In this section I argue that the switch from viewing AI explainability as improving system transparency to viewing explainability as accounting for outputs is no small move. It is a switch from adopting what Daniel Dennett has called the "design stance" towards AI-enabled expert systems to adopting what I call the "testimonial stance" towards AI-enabled expert systems. This switch has significant implications for how we think about the quality of an explanation.

The design stance is one of three intellectual strategies Dennett (1987: 16-18, 48-51) outlines for understanding and predicting the behavior of non-human external entities. The most fine-grained of the three is the physical stance according to which one understands the behavior of an entity by analyzing the physical, chemical, or biological properties or laws that govern it. For example, in adopting the physical stance toward a pendulum one might predict the pendulum's movement by referencing Newton's laws of motion. The physical stance is not, however, always the best intellectual strategy to adopt for understanding a system's behavior. To understand, for instance, all the inner workings of a car purely in terms of physical and chemical laws can be quite difficult; a car engine is more complicated than a simple pendulum. Similarly, understanding an AI system's behavior according to the physical stance would require thinking about the movement of electrons along circuits. This is not the kind of information that would help a user decide whether to believe an AI system's outputs.

Dennett's design stance zooms out from the physical stance. By adopting the design stance, one understands a system's behavior at a higher level of abstraction in terms of its construction, design purpose, and functionality. The design stance is how most people understand prototypical instruments and appraise their functionality. For example, when a car mechanic appraises a car engine from the point of view of the design stance, she asks whether all of the engine's components are working together as they were designed in order to safely transfer rotational force into the rear wheel axle. In comparison, adopting the design stance towards an AI system involves understanding the system's functionality in terms of its algorithmic construction. One would ask how the input data is

processed by the AI system in such a way that it yields the given outputs. Those who understand AI explainability merely in terms of transparency take the design stance towards AI systems like they would any other machine or instrument.

However, just as the physical stance can require a person to deal with unwieldy and unnecessary detail, so too can the design stance. Such is often the case for AI systems. As I have explained, the challenge to XAI research is precisely that looking into how a system works as one would by lifting the hood of a car is often an impracticable strategy for gleaning insight as to how well the system has performed. AI systems are too complex to expect such design-level appraisal to yield information that would help users epistemically justify their beliefs in system outputs.

Dennett's intentional stance is the third and most abstract intellectual strategy he presents for understanding artefactual systems. Adopting the intentional stance towards an external entity involves attributing to that entity intentions, beliefs, desires, or goals in order to predict system behavior.[2] For example, a chess player significantly narrows the set of her opponent's next possible moves if she attributes to her opponent (human or AI) the intention of winning. Dennett argues that the intentional stance enjoys predictive success because systems are typically well-designed (natural systems by evolution and artifactual systems by humans) to exhibit behaviors that satisfy needs and desires the system ought to have given its capacities and environmental context. For example, attributing both human drivers and autonomous vehicles with the desire not to crash is predictive of stopping before red lights because human drivers are evolutionarily disposed to avoid harmful behavior, and autonomous vehicles are designed to keep human operators safe.

While adopting the intentional stance is certainly not an infallible method of predicting system behavior, Dennett argues that the strategy is appropriate if thinking of an external actor *as if* an intentional agent proves to be predictively useful and, given a reasonable amount of effort, a person cannot use the more fine-grained views of the design or physical stance to more accurately predict behavior.

---

[2] Readers interested in artifact intentionality may also find Peter-Paul Verbeek's (2006, 2008) notions of "technologically mediated intentionality" and "cyborg intentionality" informative.

In Verbeek's view of "technologically mediated intentionality" an artifact is designed to physically embody and help realize the goals of its indisputably intentional human producers. Verbeek provides a simple example: a plastic cup 'asks' to be thrown in a landfill while a ceramic mug embodies the designer's intention that the mug be washed and re-used (2008: 362).

On the other hand, "cyborg intentionality" refers to how human user intentions are molded by the user's technologically mediated experience. While developers may hold specific intentions for how a technology is meant to be used (technologically mediated intentions), and users may hold specific intentions for how they would like to use the technology (user intentions), the resulting action may be a unique and often unforeseen manifestation of cyborg intentionality which is neither fully attributed to people in the artifact's chain of production nor to the user. For example, Verbeek describes how the intention of an ultrasound machine (to visualize a fetus) might combine with the un-mediated intentions of expectant parents (to care for their baby) to yield a new intention (to abort a fetus that displays signs of disease) (2008: 365-366). The new intention is made possible and implicitly encouraged by the technology's capabilities, yet it is not fully attributed to the technology.

What I call the *testimonial stance* is a variation on Dennett's intentional stance. The intentional stance describes the attribution of intentions, beliefs, desires, and goals to a system in order to better predict its behavior. To adopt the intentional stance toward an artifact is to take seriously the idea of the artifact as if an intentional entity. Similarly, adopting the testimonial stance involves evaluating various information sources in terms of characteristics attributed to prototypical testimonial speakers. It is to take seriously an analogy between human sources of information (testimonial speakers) and technological sources of information.[3] Though instead of predicting system behavior, as is the goal when adopting the intentional stance, the aim of adopting the testimonial stance is to understand how one might justify her belief in AI system outputs by comparing to how people justify their belief in human speaker claims.

As with the intentional stance, I propose that it is appropriate to adopt the testimonial stance towards any artifactual information source when it is useful to do so – that is, when by viewing the source as if a source of testimony one can make a more epistemically well-informed decision about whether to believe the source's outputs than by exerting a reasonable amount of effort to appraise the source's track record or by adopting the design stance. This will most likely be the case for those technologies which do not have well-established or easily accessible track records and which, like AI-enabled expert systems, are mechanistically complex and therefore epistemically opaque to users.

---

[3] Others have noted similarities between how people acquire information from other people (human testifiers) and how they acquire information from instruments like AI systems (Sosa, 2006; Gelfert, 2014: 27).

There is even some debate as to whether some artefactual sources of information might be considered testimonial speakers in their own right. Billy Wheeler (2020) argues that some instruments ought to be considered genuine testimonial speakers on the grounds that they have a genuine capacity to deceive. On the other hand, Sanford Goldberg (2012, 2017) rejects testimonial speakerhood to artifacts on the grounds that non-human entities are not subject to the same kinds of normative evaluations as human testifiers because they do not hold the same kind of capacities for sincerity, responsibility, belief, etc., that humans do. In a similar vein, Richard Moran (2006: 272-273) promptly dismisses the idea of instruments as testifiers based on the nature of their fallibility; while humans are irreparably unreliable mediators – the tendency to lie and speak carelessly is human nature – machine reliability consistently improves through iterations of research, design, and development. Elizabeth Fricker (2015: 117, 201) comments that automated auditory speech (like airport loudspeaker announcements) should be considered instances of "fake testimony". Hearers can receive knowledge through these kinds of communications that phenomenologically resemble instances of testimony if the content is true, but the knowledge source does not testify because it does not hold belief or knowledge of the content being conveyed. Here Fricker takes the capacity to hold doxastic states as a defining feature of testimonial speakers. Stephen Wright (2014: 251) also argues that "testimony comes from speakers that have doxastic states that instruments lack. Speakers know things, where instruments do not [. . .] This supports the claim that there are certain attitudes that we can coherently take in response to testimony that we cannot coherently take in response to an instrument." In yet a third approach, Ori Freiman and Boaz Miller (2020) take a middle ground arguing that some instruments might instead qualify as sources of "quasi-testimony". Their position is reminiscent on Don Ihde's (1990) notion of the technologies to which human users bear some relationship as being "quasi-other"; people project human properties on technologies (e.g. intelligence, honesty, intentionality) and often relate to technologies as they would other humans (e.g. by caring for the technology), but a technology can never present like another person in the fullest sense because unlike in human-human relationships, in human-technology relationships a person could never imagine herself in the technology's shoes.

I have not, however, found entering the fray to define testimony and testimonial speakerhood to be a productive strategy for investigating the justificatory value of AI explanations. We need not first qualify AI systems as testimonial speakers to find thinking about them *as if* testimonial speakers to yield useful insights. It is for this reason that I put forward the testimonial stance in analogy to Dennett's intentional stance as a way of thinking about AI systems *as if* sources of testimony when there is merit to doing so.

For simpler instruments, like microscopes or wristwatches, and for users familiar with those instruments' construction and functionality, the design stance remains accessible and will likely facilitate more accurate appraisals of system reliability.

I should note that John Zerilli (2022) has recently proposed that Dennett's intentional stance may provide a framework for understanding how AI users might assess whether an AI output or decision is justified. His discussion specifically pertains to how the intentional structure of AI explanation can allow users to employ the tools of folk psychology to justify their belief in AI outputs despite the necessarily incomplete and unfaithful (to the computational processes by which an output was derived) nature of interpretable AI explanations. My discussion of the testimonial stance is broader than Zerilli's application of the intentional stance to understanding explanations. As noted previously, the testimonial stance is a variation on the intentional stance which involves viewing an information source as if a source of testimony. Because prototypical testimonial speakers are intentional beings, interacting with and appraising AI systems as if intentional entities (entities that form beliefs and act in response to those beliefs) is one way of adopting the testimonial stance. But there is much more to the epistemology of testimony than appraising speaker intentions.

Different veins of epistemology of testimony draw attention to different factors that can influence justified belief in speaker claims such as speaker accountability and responsibility (Moran 2006), speaker sociocultural context (Fricker 2006; Lehrer 2006), a speaker's sincerity and intention to deceive or manipulate (Coady, 1992; Lackey 2006; O'Neill 2007; Wheeler 2020), a speaker's ability to defend or explain her claims (Goldman, 2001), or speaker competence and expertise (Anderson 2011; Goldman 2001).[4] Accordingly, adopting the testimonial stance towards an AI system can involve evaluations of any similarities between acquiring information from the system and from a human speaker. It may involve, as Zerilli proposes, scrutinizing the intentional structure of AI explanations. But adopting the testimonial stance may also entail attending to the norms of communication to which the technology responds (or seems to respond), or it might prompt investigation into various environmental factors that influence the technological speaker's trustworthiness, like the norms, motivations, or intentions that influence the technology's producers or employers. I touch on many of these factors in Part II of this dissertation. In this and the following chapter I am, like Zerilli, specifically interested in AI explanations. I will, however, discuss on a more general level how XAI researchers and AI users adopt the testimonial stance towards AI systems with respect to how they think about AI explanation form and function.

In terms of function, AI explanations are viewed as a means to some practical end, here as a vehicle for helping users justify their belief in AI outputs in much the same way as human expert explanations help novices decide whether to believe expert claims. I draw attention to two main parallels in the form of expert and AI explanations.

[4] Axel Gelfert (2014) provides a thorough overview of the literature and a useful reading guide.

First, both AI and human expert explanations are *post hoc*. There is no direct window to the real-time inner workings of the human mind nor do AI explanations provide a direct window to the inner workings of an AI system like a car owner might acquire by lifting the hood of her car. As such, AI explanations are better thought of like a report on the state of a car's engine as provided by a mechanic to a naïve car owner (one who knows very little about cars) after the fact. The mechanic looks under the hood, appraises the engine, identifies any issues, and draws conclusions about necessary repairs. She then offers an explanation to support her conclusions, and the car owner must decide whether to believe the mechanic's claim that the repairs are indeed necessary. The challenge posed by post hoc explanations is that there is a gap between the initial reasoning or computational process by which an expert or AI system derives a conclusion and the report offered thereof. In the gap between the initial reasoning process and the report, some determination must be made about what information an explanation recipient is to receive about the original reasoning process and what information is to be withheld, simplified, rearticulated, or fabricated in the name of recipient understanding.[5] In contrast, direct observation of internal mechanisms involves no such gap. All determinations about what information is relevant to an appraisal of the system's output are made by the observer, and no interpretation of that information is made on the observer's behalf. For example, if the car owner were to pop the hood of her car herself, all the visual, auditory, and perhaps olfactory information she would receive about her car's engine would be unmediated by an external entity. It is up to her to use what knowledge she has about cars to sort out what information is relevant to the diagnosis of her car troubles and what is not.

This leads to the second parallel. Both AI and expert explanations are *partial*; they most often do not provide full, unaltered descriptions of the internal reasoning/computational processes employed to derive a conclusion. In the case of explanations offered by human experts to a novice, for example, an expert delivers altered accounts of her reasoning because a modified or alternative account would be more cognitively accessible to the novice. The expert utilizes a body of knowledge and cognitive skills that are, by definition, not held by a novice, and therefore any account of her reasoning that she offers to the novice must be rearticulated and tailored to novice understanding. Never entertained is the idea that the expert should instead simplify the way in which she reasons to a conclusion. Indeed, we hope that the experts we consult employ the full extent of their knowledge, training, and skill in providing us with answers and advice. Similarly, in switching from a design stance to a testimonial stance, the idea of simplifying the AI system itself is taken off the table. If the full beneficial potential of AI technologies is to be realized, then we must be willing to build increasingly powerful, and

[5] I take inspiration from John Searle's (2001: 13-17, 62) discussion of gaps in human reasoning and decision-making processes. There is a necessary gap between a person's beliefs and desires because beliefs and desires are not, on their own, sufficient to cause a decision to be made. There is a gap between decision and action. The decision is not causally sufficient for action. For example, you may decide to go to the polls to vote, but for some reason (or, perhaps, for no reason at all) you do not. Third, there is a gap between the initiation of an action and the completion of the action. Initialization is usually not causally sufficient for completion; some continuous effort is required to keep the process moving along.

therefore increasingly complex systems. Consequently, there is a doing-saying mismatch (to borrow a term from Jutta Schickore, 2008) between an AI system's internal process and its report thereof; what an AI system 'says' it does and what it 'does' do to derive an output will always differ to some degree. It is the same kind of doing-saying mismatch that we see in the case of expert explanations.

The switch from viewing AI explanations through the lens of the design stance – where explainability is understood in terms of mere transparency – to the testimonial stance – where explanations are acknowledged as post hoc and partial account of reasoning – raises questions about how and to what extent explanatory accounts can and should be altered (simplified or otherwise manipulated) in the name of understandability. An important challenge for XAI researchers and epistemologists of testimony alike is understanding the degree to which an explanation's abstraction from an original reasoning/computation process influences the justificatory value of the explanation either for the worse or for the better. The upswing of the testimonial stance is that if there is potential for human expert and AI explanations to be justificatorally valuable despite, or indeed, because of, abstractions, then the explainability-accuracy trade-off may fall away; there may be cases in which AI systems may retain complexity (and associated power) while still providing understandable explanations that can help users to epistemically justify their belief in the system outputs. There is no such opportunity where AI explainability is understood merely in terms of transparency.

In the remainder of this chapter, I investigate what factors influence the justificatory value of the kinds of post hoc and partial explanations offered by human experts and AI systems. The following section begins by considering the well-established philosophical literature on scientific explanation as a starting point for the investigation.

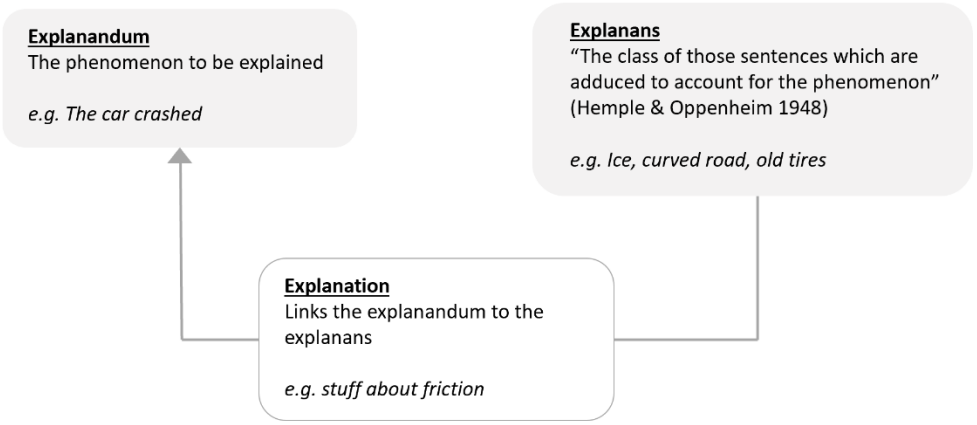
2. PHILOSOPHY OF EXPLANATION

For those looking to better understand what makes for a good AI explanation, the literature on philosophy of scientific explanation is an obvious place to begin. Philosophers of explanation have engaged with questions of explanation structure and quality since the Socratic era with more modern conversations building on Hempel (1965) and Hempel and Oppenheim (1948). It makes sense that philosophers looking to investigate AI explainability would turn to philosophy of explanation for guidance. For example, Adrian Erasmus et. al. (2021) leverage classic philosophical theories of explanation to describe precisely what it means for an AI system to be "explainable". Christopher Grimsley (2021) argues that literature on the philosophy of scientific explanation is "uniquely equipped" to guide discussion of AI explainability. Rune Nyrup and Diana Robinson (2022) use pragmatist accounts of explanation to investigate requirements for explainable AI in medical settings, and Keiron O'Hara (2020) understands explanation from the point of view of philosophy of science in order to elucidate legal requirements for AI explainability set out by the General Data Protection Regulation (GDPR).

However, while I agree that the philosophy of explanation has much to offer philosophical discussions of explainable AI, I caution that the vast body of classic literature in philosophy of explanation is not all well suited to a discussion on how AI explanations help a person evaluate whether they ought to believe a particular output. It is easy to spiral down an unproductive rabbit hole. Troubles arise because the notion of explanation we have in mind when discussing explainable AI – as akin to the post hoc and partial accounts of reasoning offered by human experts – differs in several key ways from how explanations are most commonly understood by philosophers of explanation. In what follows I first describe what I call the *paradigm model of explanation* within philosophy of science, and then I expound on those key differences. In light of the discussions contained in this and the previous section, in Section 3 I propose a framework to guide our thinking about the justificatory value of AI and expert explanations alike.

2.1. The paradigm model of explanation

Approaches adopted by philosophers of science to understand explanation vary, though most follow a general structure. Within the paradigm model of explanation, explanations are generally considered to provide answers to "why" questions which ask why things happen or why things are the way they are by linking the *explanandum* – "the phenomenon to be explained" – to *explanans* – "the class of those sentences which are adduced to account for the phenomenon" (Hempel and Oppenheim, 1948). For example, a question might pose the explanandum "Why did the car stall?" to which another might provide an answer "because it has no fuel". 'That the car stalled' is the explanandum and 'no fuel' is the explanans. The *process of explanation* links the two, for instance, by providing some statement (or series of statements) to the effect that cars require fuel to run.



[Figure 1.1] The general structure of an explanation

A good explanation is one that accounts well for the observed phenomenon. There are various metrics by which philosophers determine how well an explanation accounts for a phenomenon. For instance, one might look at how predictive the explanation is of the phenomenon being explained (Hempel & Oppenheim, 1948). A good theory of planetary motion, for example, should be predictive of the location of the planets far into the future. Others might appraise whether the explanation logically builds from natural laws or law-like premises (Salmon, 1984; Elgin & Sober, 2002), or look to whether the explanation adds to or detracts from the coherence of existing networks of belief and knowledge (Thagard, 1989; Glass, 2007). What these accounts have in common is that they generally hold that a good explanation will meet some kind of veridicality condition in which its premises (explanans) are either true or "well-confirmed" and stand in the right kind of relationship to the explananda (*inter alia* see Achinstein (2010), Hempel (1965: 248), and Strevens (2008)).

Furthermore, most philosophers of science are non-cognitivists about explanation; the quality of an explanation depends only upon how well the explanation accounts for the explanandum while how understandable the explanation is to any particular person is beside the point (Potochnik, 2016; Strevens, 2008; Craver, 2014). As Michael Strevens (2008) writes, an explanation "exists before science; it is what scientists, when they arrive, undertake to discover" and once discovered, adds to the stock of scientific knowledge about the world (40). If explanation quality were to have a cognitive dimension, the implication would be that scientific discoveries would be undermined by the ability of the most cognitively limited humans to understand those accounts. But there is something intuitively wrong here. One person's inability to grasp a scientific explanation for some phenomenon, is not an argument against it. A good explanation accounting for the existence of black holes, for instance, is still a good explanation even if only the Stephen Hawkings and Kip Thornes of the world can make heads or tails of it.

In summary, according to the paradigm model of explanation, an explanation describes some relationship between an explanans and explanandum. A good explanation must meet some veridicality condition, and whether it is comprehensible to a recipient is beside the point. There are of course variations on this model, and I will discuss them shortly, but for the time being the paradigm model of explanation provides a starting point for thinking about how AI explainability fits within broader discussions of philosophy of explanation. For instance, Adrian Erasmus et al. (2021) demonstrate that there is a straightforward way in which AI explainability can be thought of in these standard philosophical terms. The explanandum is the system output (some statement, prediction, or recommendation). The explanans include data inputted to the system, features of the algorithm's architecture, statistical weights assigned to different nodes in an artificial neural network (a common AI architecture), or the flow of electrons in a circuit.

With these explananda and explanans in mind, Erasmus et al. describe how artificial neural networks (ANNs) are explainable according to four standard philosophical accounts of explanation that outline different ways in which explanans and explananda can be related. They begin with the

Deductive-Nomological (DN) theory of explanation outlined by Hempel and Oppenheim (1948). In the DN model, explanans are connected to explanandum via a process of deduction in which the explanandum is the logical consequence of the explanans and the explanans include at least on nomic premise such as a law of nature or a law-like premise. Erasmus et. al. use the example of a cancer diagnostic medical AI system (MAIS) to describe how to construct a DN explanation for the outputs of a ResNet-18 convolutional neural network – a neural network into which images of breast scans are inputted and then processed through 17 convolution layers of weighted nodes and edges to output a tissue classification (Lehman et. al. 2019). They write:

> A DN explanation of how the MAIS assesses an input image involves listing the weights attached to each and every node and edge at every convolution stage and the weights of the fully connected network along with the assigned numerical values being fed into the input layer and the network architecture. Once we have that, we can list the values for the classifications the MAIS learned in the training and testing phases of development, and see that its classification of the image is based on comparing the ranges of these classifications with the output value of the image. In doing so, we are explaining the explanandum – here, the MAIS classifying of image *I* as classification *c* – using an explanans consisting of a law-like premises – in this case, how the weights of all relevant nodes and edges produced the output value, along with the law that an output is assigned to the most probable class – and additional information about *I* – which includes the set of input values assigned to *I*, and the output value *c* (Erasmus et. al. 2021, 12-13).

Erasmus et al. provide a similar analysis with respect to the Inductive-Statistical (IS) model of scientific explanation (Hempel, 1965), in which an explanandum is inferred from statistical laws and empirical information, as well as for Wesley Salmon's (1984) Causal-Mechanical (CM) model, in which an explanandum is situated within the causal structure of the world by showing that the explanandum results from a series of causal process and interactions, and for the New-Mechanist (NM) model (Craver & Darden 2013; Machamer et al., 2000; Bechtel, 2011) in which an explanation identifies a series of activities and events that lead to a phenomenon as well the starting and termination conditions for the processes. It is Erasmus et al.'s goal to show that there is no need to reinvent the wheel, no need to begin explicating explanation anew because artificial intelligence has entered the ring. According to standard notions of explanation developed by philosophers of science over the past century there exists some logical, statistical, or causal way to account for AI outputs regardless of system complexity, and as such, Erasmus et al. note, there is no explainability-accuracy tradeoff; regardless of how complex an AI system is, it is possible to produce a veridical account for the explanandum "the AI system outputted X".

I agree that there is no need to reinvent the wheel, and I accept that there is no explainability-accuracy trade-off if we understand explanations accounting to the paradigm model discussed by philosophers of science. However, Erasmus's account illustrates exactly why we need to be careful about how discussions regarding AI explainability are situated within the philosophy of explanation literature. Theirs is a perfect example of the kind of rabbit hole misadventure we wish to avoid. As prefaced in the introduction, AI explanations are supposed to help overcome challenges posed by AI system opacity in order to achieve some further end. In the context of this chapter, that

end is to provide users with the kind of information they can use to help justify their belief in AI outputs and recommendations. However, in thinking about AI explainability in line with the paradigm model of explanation, Erasmus et al. miss the point. They may, for example, show that some explanation for an AI output can be produced regardless of system complexity and that in this way the accuracy-explainability trade-off falls apart, but what use is this observation? The pragmatic challenges posed by AI opacity remain unaddressed. For instance, even if for a particular ANN the DN explanation articulated above is filled out with the weights of all relevant nodes and edges, the explanation would still not help the typical user make an epistemically well-informed decision about whether to believe the system's outputs.

There is growing recognition that the pragmatic ends of explanation are of central importance to any discussion about the quality and construction of AI explanations (Krishnan, 2020; Páez, 2020; Nyrup & Robinson, 2022). Andrés Páez calls this the "pragmatic turn" in XAI literature; depending on which pragmatic end is in focus (safety, accountability, justification, etc.) the characteristics of a desirable explanation will likely differ. Intended function must be taken into account when describing requirements for good explanations or appraising explanation quality. AI explanations are not, as Erasmus et. al. see them, ends in and of themselves.

None of this is to say philosophy of explanation is wholly ill-suited to a discussion of explainable AI. To the contrary, consider Rune Nyrup and Diana Robinson's (2022) discussion of medical XAI. Nyrup and Robinson identify "enabling informed consent" as a key pragmatic end for explanations offered by AI systems employed in health care settings. Patient understanding is generally thought necessary for informed consent to obtain. As such, Nyrup and Robinson fruitfully inform their analysis of desirable medical XAI qualities by drawing on discussions about scientific explanations formulated to promote recipient understanding of scientific theories and phenomena. So, my point is not that philosophy of explanation is ill-suited to discussion of XAI. Nyrup and Robinson show that it can be very well suited. My point is only that we need to be careful about how AI explainability is couched in the philosophy of explanation literature. I show how the kinds of post hoc and partial AI and expert explanations we care about in this chapter stray from the paradigm model, and in so doing I point to areas of explanation literature more well suited to the present discussion about the justificatory value of explanations in particular. My overarching goal is to orient AI explanations within the philosophy of explanation literature in such a way as to encourage pragmatically oriented conversations (like Nyrup and Robinsons' and my own), and to prevent tangents like Erasmus et al.'s.

2.2. Essential cognitive dimension

In the paradigm case of explanation, an explanation recipient's ability to understand an explanation is not typically considered a necessary characteristic of a good explanation. Most

philosophers of explanation are non-cognitivists about explanation because they understand an explanation as something that exists out in the world, a set of facts waiting to be discovered that accounts for why the world is the way it is. Whether any particular person can make sense of those facts is irrelevant.

However, for the sake of discussing the justificatory value of AI and expert explanations, the non-cognitivist view of explanation is incomplete. If AI and expert explanations are means to help achieve some end, and not an end themselves, then the quality of an explanation cannot solely be dependent upon how well the explanation accounts for a phenomenon. Put otherwise, the quality of an explanation cannot be evaluated in isolation from an explanation recipient's ability to use the content of the explanation to help achieve the desired end.

Thankfully, while most philosophers of explanation take as fundamental the notion of explanations as set of facts that account for a phenomenon (the paradigm understanding), Michael Strevens (2008, 2016) points out that there is a pragmatist subset of those philosophers who also emphasize the role of explanations as communicative acts – as means of communicating those explanatory accounts to others. These philosophers defend the view that explanations are communications between two parties, a sender and receiver, and therefore an explanation is only successful if the recipient can understand the information with which she is provided. Cognitivists about explanation argue that there *is* a necessary cognitive dimension to good explanations.

Perhaps most notable among the cognitivists is Peter Achinstein (1983, 2010) who describes an explanation as an "illocutionary act" in which an explainer's intention to convey understanding to her audience is an essential component of a good explanation. Achinstein (2010: Ch.7) differentiates between "correct" explanations which satisfy a veridicality condition and "good" explanations which are correct as well as sensitive to an explanation recipient's ability to understand an explanation (as restricted by time, background knowledge, and cognitive capacity) and responsive to the explanation recipient's interests and goals. For example, with respect to AI, a developer might seek an explanation for an output to identify errors in the system design or to improve system performance, or a user might seek an explanation to justify her belief in the system's outputs or to learn from the system's reasoning process. For instance, an AI system used to aid in drug discovery might uncover new laws of chemistry or physics. Users and developers will have different capacities for understanding, and their different interests will be satisfied by different explanatory content. In this way, it is possible for an explanation to be correct – in that it elucidates explanatory connections between true premises – but not good if it does not cater to audience interests and understanding. For example, the DN explanation Erasmus et al. provided (see excerpt two pages back) for why the MAIS produced a particular output may be true, but listing the statistical weights of the interconnected nodes and edges that comprise a deep neural network would not provide a lay user with information she can comprehend so as to inform her decision about whether to believe the system's output. Maya Krishnan

(2020) makes a similar observation about Causal Mechanical explanations offered for ANN outputs. Krishnan writes:

> For instance, if you ask a person why they have given an answer to a particular question and they respond with an account of how their neurons are firing, they have given you information about the causal process that subserves the generation of their answer, without telling you anything that has clear significance to the justification of their answer... there remains a difficult task of translating information about causal processes into considerations relevant to the justification of a categorization (494).

On the topic of catering to audience interests and needs, Bas van Fraassen (1980) emphasizes the importance of context sensitivity to the interpretation of why questions (See also de Regt, 2009a, 2009b). Van Fraassen points out that when "why" questions are asked, the inquirer implies that she would like to understand why one explanandum was offered instead of a contrasting explanandum. For example, the question "Why are the leaves orange?" might ask why the leaves on this tree are orange while the leaves on that other tree are purple, or the inquirer might want to know why the leaves of this tree are orange when at a previous point in time they were green. The appropriate interpretation of the question depends on the context, and a different explanation will be good in each case. Again, we see parallel concerns expressed in relation to explainable AI. For example, an explanation for why a loan appraisal system denied a loan to a customer could list all factors used in the decision, identify the most influential factor accounting for the rejection, describe the closest possible world in which the loan would have been granted, or list the changes the subject could realistically make to achieve a different outcome in the future. Which explanation is good will depend on who is seeking the explanation (AI developer, banker, or customer) and for what purpose the explanation is being requested (Watcher et. al. 2018; Zednik, 2019; Krishnan, 2020).

In this way, cognitivists and non-cognitivists about explanation are often framed in opposition to one another such that understandability and context sensitivity either are or are not essential characteristics of good explanations. However, I posit that there is no such conflict. Rather, the two parties engage with two independent notions of explanation: *explanations-as-accounts*, for which only the veridicality of the explanation is considered, and *explanations-as-communications* for which the cognitive dimensions of understanding and context sensitivity must also be taken into consideration. Where disagreement arises around the necessity of a cognitive dimension to explanations, it is only a perceived conflict resulting from confused terminology where the term "explanation" is used to refer both to explanations-as-accounts and to explanations-as-communications. Therefore, when engaging in a philosophical conversation about explanations we need not decide which notion of explanation is correct, but only to identify which explanation type we have in mind. Erasmus et al., however, gets lost in the confusion. Led by the primary emphasis on explanations-as-accounts within philosophy of explanation literature, Erasmus et. al. oriented their discussion of AI explainability according to that notion. However, the stated aim of their discussion is to expound on a practical problem – the accuracy-explainability trade-off – to which explanation has been proposed as a solution, but where

explanation is clearly meant to be understood as a kind of communication offered to the user by the AI system.[6] Consequently, while I believe Erasmus et al. provide an excellent account of how AI explainability can be understood in terms of non-cognitive theories of explanation, in adopting the non-cognitive perspective they miss the point entirely.

To conclude this subsection, I posit that there is still an open question about whether a good explanation-as-communication must be a good explanation-as-account as well. The question hinges on whether explanations-as-communication must always be veridical. In the course of the next two subsections, I propose that the veridicality condition may not need to be satisfied when the primary goal of the explanation is providing justification for a claim.

2.3. Which explanandum?

In order to discuss explanation veridicality, it must first be made clear what, exactly, is being explained. However, in the case of explanations offered in support of claims, it is not immediately obvious what the explanandum is. When explanations are understood in the traditional sense as accounting for some phenomenon (or communicating accounts thereof), the explanandum is a given. It is some observation or prediction about the world – e.g. the car stalled, the leaves turned orange, tomorrow it will rain – and the goal is to learn, discover, understand, or describe why the explanandum is or is likely to be. However, when talking about AI explainability, there are two distinct explananda one might have in mind. Any productive discussion about the quality of AI explanations must first identify which is in focus to avoid confusion.

First, the explanandum could be the event in which an AI system provides an output. That is, when an AI system produces an output X, the explanandum is "the AI system outputted X". Those who take the design stance towards AI systems have this explanandum in mind; to account for the event in which the AI system outputted X by elucidating how the AI system derived X.

However, in adopting the testimonial stance towards an AI system, we are primarily interested in a second possible explanandum: the output itself. Instead of asking "Why did the AI system output X" or "Why did the physician recommend Y?", our primary interest is in asking "Why X/Y?" or more specifically, "Should X/Y be believed?".

The distinction is important because there are various good (understandable and true) explanations for the explananda "the AI system outputted X" or "the physician recommended Y" that do not necessarily provide an explanation recipient with the kind of information she needs to help her decide whether to believe X or Y. For example, a correct and understandable causal-mechanical account might describe how the vibrations of the physician's vocal cords resulted in a specific sequence of sound waves which accounts for her utterance, "I recommend treatment Y". Similarly, causal mechanical accounts of ANN outputs would cite various causal interactions among the signal

[6] Kieron O'Hara (2020) offers an extended argument for considering AI explanations to be the kinds of "illocutionary acts" described by Achinstein.

pathways between weighted nodes and edges that led to output X. Such explanations may be correct and understandable accounts for the given explananda, but they fail to provide any reason to believe the content of the AI system's output or expert's claim. So, going forward we must keep the content of the expert's/AI's outputs as our main focus.

2.4. Nonessential veridicality

Having adopted the testimonial stance, we are viewing explanations as communications which must be understandable and context sensitive and which are primarily meant to account for the content of a claim/output (X), not the event in which that claim/output was offered (the AI system outputted X). Now I discuss whether an expert's/AI's accounting for X must be veridical to help a recipient justify her (dis)belief in the output/claim.

According to the paradigm model of explanation, a good explanation must satisfy some veridicality condition for explanatory correctness: a good explanation will be true or otherwise well-evidenced to be an accurate representation of reality. This is the case whether explanations are understood as accounts or as communications (of accounts) for which understandability and sensitivity to recipient interests is required. However, when explanations are offered to provide justification for an expert claim or an AI output, we need to be careful about how we discuss veridicality. There are two senses of veridicality at play – that is, two dimensions along which an explanation offered by an AI/expert can be judged as true.

The first sense of veridicality (V1) is as traditionally understood by philosophers of science. If an AI system or expert makes claim X (e.g. Amoxicillin will cure the patient's staphylococcus infection), then a veridical explanation is one in which the premises are true or well-confirmed and stand in the right kind of relationship to X such that they account for X. Examples of how such explanations might be structured include the aforementioned theories of explanation proposed by philosophers of science. For example, an inductive statistical (IS) explanation for X would involve reference to supposed statistical laws such as the probability of patient recovery from staphylococcus infections given the administration of Amoxicillin. Alternatively, a Causal Mechanical (CM) explanation for X might reference true causes of the infection (e.g. overwhelming levels of staphylococcus bacteria) and describe true facts about how the antibiotic intervenes on the mechanisms of infection (i.e. how the Amoxicillin antibiotic kills staphylococcus bacteria or prevents them from replicating).

The second sense of veridicality (V2) is about an explanation's fidelity to the expert's or AI system's original reasoning or derivation process for X. V2-veridically still refers to whether an explanation is an accurate representation of reality (the way the world is) in the sense that there exists a specific way in which an expert reasoned to a conclusion or an AI system derived an output. The expert/AI can recount that process more or less accurately.

Let me quickly put out a small fire. To be concerned with the V2-veridicality of AI explanations might seem quite a bit like adopting the design stance towards AI systems; by adopting the design stance one is primarily interested in seeing how the AI system works, and V2-veridicality provides a glimpse of AI system functionality by recounting how an output was produced. However, while the design stance and V2-veridicallity certainly touch on a similar theme, I would like to emphasize that they are not equivalent. By adopting the design stance, one is interested in system transparency and in the explanandum "the AI system outputted X". One is interested in appraising the inner workings of the system, understanding how the system works, and how it produces its outputs. V2-veridicality, on the other hand, is a characteristic of an explanation offered in support of an output, not a characteristic, like transparency, of the entity producing the output. To be concerned with V2-veridicality is to be interested not just in how an output or claim was derived by an AI system or human expert, but in how it/she *says* the output was derived, and, as prefaced in Section 1, this kind of doing-saying mismatch may have important implications for how much the explanation helps a recipient decide whether to believe output (I tackle this topic in the next section). Therefore, V2-veridicality is a concern only under the testimonial stance and we can discuss V2-veridicality while keeping the explanandum "X" in focus. Now back to the topic at hand.:

It is important to distinguish between V1- and V2-veridicality because they are two independent qualities that can be used to describe a single explanation. For instance, if a physician reasons poorly to a diagnosis in that she bases her reasoning in one or more false premises (e.g. perhaps she misremembers a relevant biological mechanism) and then honestly communicates her reasoning process to a patient, then the explanation is V2-veridical (it is an accurate representation of her reasoning process) but non-veridical in the sense of V1 (that reasoning processes is not a good explanation for the diagnosis). Inversely, an explanation can be V1 veridical (e.g. a physician accurately describes the biological mechanism underpinning a set of symptoms) but non-V2-veridical (the physician's reference to biological mechanism was a post hoc rationalization of her diagnosis which was originally based on intuition).

If we control for V2 for the moment (i.e. assume an explanation is V2-veridical), note that an explanation's justificatory value does not hinge on its V1-veridicality. An explanation is justificatorally valuable insofar as it contains information that can help a person decide whether to believe or disbelieve a claim or output by appraising the explanation for coherence with her own background knowledge. Accordingly, an explanation can be justificatorally valuable when it is not V1-veridical because the very fact that it is not V1-veridical can give a person reason to think that a claim is not likely to be true. For instance, non-V1-veridical explanations might contain illogical gaps in reasoning or false premises which (assuming the recipient holds true or epistemically well-founded

background beliefs) would appropriately lead a perceptive recipient to disbelieve the claim the explanation is meant to support.[7] I expand on mechanisms of justification in the next section.

So, now setting aside V1, the subsequent question is about the necessity of V2-veridicality to justificatory value. At first look, it seems that V2 veridicality must be of some importance. If a person or AI system does not faithfully recount her/its internal reasoning or derivation process, then the explanation recipient cannot meaningfully appraise the explanation for illogical steps or false premises. That said, it is easy to think of examples in which non-V2-veridical explanations might be preferable. For instance, if a chiropractor bases her diagnosis of a patient's back pain on intuition, an explanation along the lines of "It felt right in my gut" provides no information the patient could use to help her decide whether to believe or disbelieve the diagnosis. In such a case, a non-V2-veridical explanation that presents a post hoc justification for the chiropractor's gut reaction would be more informative. Lacking any information about the chiropractor's track record for true diagnoses, if the chiropractor backs her instinct with seemingly non-V1-veridical justifications, then one would have some reason to disbelieve the chiropractor's claim.

In the following section I discuss in greater detail how abstractions from the original reasoning/derivation process affect the justificatory value of an expert/AI explanation. But before I continue, it will help to have a more clear idea of what I mean by V2-veridical explanations as "faithfully representing" an original reasoning/derivation process.

There is an area of philosophy of scientific explanation which can help inform this discussion. The literature on explanatory models deals almost exclusively with issues of 'abstraction from the original'. There are many kinds of models discussed within the literature on scientific models and model making including physical or "material" models, simulations, analogies, and equations. However, all models can generally be understood as epistemic representations of some object, system, or phenomenon of interest (Bolinska, 2013). The idea of an epistemic representation is that by holding key similarities to the object of interest a model can be studied, experiment on, or used as a pedagogical tool in place of the original which may be too large, too complex, too far future or past, etc. for a person to study or experiment on directly (Morrison, 2009). The obvious debate that ensues is about how and to what extent a model can be simplified or otherwise altered from the target system such that it still serves its intended purpose well.

As post hoc and partial accounts of reasoning, expert and AI explanations can be thought of like models of the original reasoning process or computational process by which experts and AI systems derive their conclusions. Simplified and tailored accounts of the original reasoning/computational processes must be offered for the sake of recipient comprehension and to respond to recipient interests.

[7] See Chapter 2 for further discussion about this assumption

There are two kinds of 'abstractions from the original' discussed in the scientific models literature. The first, and most discussed, are idealization. Idealizations, simply put, are abstractions in detail or simplifications used to make models more understandable to and manageable by the users (*Inter alia* see Jebeile & Kennedy, 2015; McMullin, 1968, 1985; Friedman, 1974). Afterall, as Sandra Mitchell and Angela Gronenborn (2017: 706) write, "by meeting a strong standard for completeness, a description or model would fail to be a representation; it would be a duplicate. For the purposes of facilitating explanation and prediction, it would be no better than engaging directly with the very system we are trying to understand".

In terms of detail, an account of reasoning can also be reported in full or simplified by stripping it of detail in much the same way as one might idealize an explanatory model or account to make it cognitively accessible while maintaining explanatory correctness. For example, Michael Strevens' (2008: Ch.4) "Kairetic account" of explanation – explanation is here understood in the traditional sense held by philosophers of science as 'that which account for a phenomena' – recommends the production of an understandable explanation via a process of idealization in which all irrelevant factors to a phenomena's occurrence (also called non-difference makers) are stripped from an explanation such that all that remains are difference makers. Difference makers are key premises in an explanation of an event or phenomenon that, if changed or omitted, would explain a different observation or would no longer logically lead to a conclusion. From this minimal state Strevens describes how an explanation can be deepened, elongated, and intensified to cater to specific explainee interests and to match the explanation recipient's capacity to understand while ensuring the explanation retains its veridicality. Here Strevens refers to what I have called V1-verdicality; the explanation is still comprised of true or well-founded premises (or in the case of explanatory models, features) that stand in the correct relation to one another such that they account for the explanandum. For example, the explanation "the car spun out of control because it slipped on the ice" is true despite lacking details about friction coefficients and horizontal forces on a vehicle taking a corner. But we can see how the same process of idealization can also be used to simplify an explanation – here understood as a post hoc account of reasoning – without reducing its V2-veridicaltiy. For example, imagine a crime scene investigator reasoned that a hunting arrow was shot from the third story window, second from the left, of the building across the street. She came to her conclusion via a process of trigonometric calculations accounting for wind speed and the average draw weight of an adult compound bow. To the press, she simply explains that she calculated the arrow's origin based on its entry angle to the crime scene. Regardless of whether the investigator's conclusion is true or her reasoning process sound – that is, regardless of the V1-veridicallity of the explanation she provides – it is V2-veridical; the explanation does not distort the investigator's reasoning process. It only leaves out detail.

The second way to abstract a model from the original target system is via fictionalization. As described by Alisa Bokulich (2011), fictionalization differs from idealization in that a fictionalized

model cannot be made identical to the modeled systems via a "de-idealization analysis" – that is, by adding back in omitted details. Fictionalized models incorporate features that are patently false. Bokulich describes, for example, Niels Bohr's model of a hydrogen atom as one such fictionalized model. In Bohr's model the hydrogen atom's single electron gains or loses energy by jumping from one stationary state (also called an orbit) to another. Within a given stationary state, the energy of the electron is constant. Today, however, we know that Bohr's orbits are fictions; modern quantum mechanics shows that an electron does not exist in a single orbit at any one time, but that the electron is better described as occupying a cloud of probability density around the atom's nucleus.

AI and expert explanations can also incorporate fictionalizations of original derivation/reasoning processes. Such is the case where analogies and examples are introduced or motivations and thought processes are cited that were not originally used by the explainer to derive her/its conclusion. For instance, an AI system might be optimized to produce psychologically satisfying explanations by mimicking the kinds of explanations offered by human experts (Harrison et al., 2017). This mimicry comes at the expense of accurate reporting on the processes employed by the AI system to derive its conclusion. Christopher Grimsley (2021) calls such fictionalized accounts of reasoning produced by AI systems "how possibly" as opposed to "how actually" explanations. Going forward I will refer to fictionalizations introduced to post hoc accounts of reasoning as reductions in explanation "reflectivity", where reflectivity refers to an explanation's fidelity to an original reasoning process. An explanation is less faithful to the original reasoning – that is, less V2-veridical with respect to the original reasoning process – the more it distorts the original reasoning/computational processes via the kinds of fictionalizations described above.

There is much ongoing debate as to whether model idealization and fictionalizations affect a model's explanatory power, that is, the extent to which it does indeed account for a phenomenon (Bokulich, 2009, 2011, 2016; Cartwright, 1983; de Regt and Gijsbers, 2016). But I am not going to engage too deeply with the debate here lest I wander down one of those rabbit holes I caution we avoid. The scientific models literature is large and enjoys very little consensus especially on questions of fictionalization.[8] But more importantly, for our present purposes, the concepts of idealization and

[8] In line with the paradigm model of explanation, most philosophers are realists about explanatory correctness and defend the idea that a model is explanatory despite its idealizations; at best idealizations have a neutral effect on explanatory power but otherwise detract from it (Jebeile & Kennedy, 2015). For example, Michael Strevens (2008) and Ernan McMullin (1968, 1985) argue that since idealized models only leave out factors irrelevant to their explanatory task, the idealizations are neutral with respect to the model's explanatory power. Michael Friedman (1974), on the other hand, argues that in accordance with the unification account of explanation, the more realistic (less idealized) a model is, the more explanatory it is. What these philosophers have in common, however, is the belief that a model's explanatory power is rooted in its similarity to the target system it represents.

There is, however, another school of thought that holds that a model's idealizations and distortions can contribute to a model's explanatory power. For example, Nancy Cartwright (1983) holds that "the truth doesn't explain much" and in her simulacrum account of explanation argues that there is no de facto tradeoff between a model's representational accuracy and its explanatory power – the two functions should be kept distinct. Also see Bokulich (2009, 2011, 2016) on genuinely explanatory fictionalizations.

fictionalization alone are useful tools for discussing how an explanation (a post hoc account of reasoning) can be abstracted in detail and reflectivity to meet recipient needs. The next section investigates how idealizations and fictionalizations contribute to or detract from an explanation's ability to provide recipients with information they can use to help justify their beliefs in claims and outputs. But first, because my discussion so far has been complicated, a quick summary:

–

When mounting a philosophical discussion about AI explainability, an obvious place to start is with philosophy of explanation. After all, both philosophers of explanation and XAI researchers consider explanations to be important, and the former has long made a study of explanation characteristics and quality. However, in this section I have shown that the philosophy of explanation literature can easily mislead investigations into the quality of AI and expert explanations if not navigated carefully.

The paradigm understanding of explanation held by most philosophers of explanation is of explanations-as-accounts; an explanation is something that exists out in the world, a set of facts about why something is the way it is, waiting to be discovered. As such, an explanation's quality refers primarily to its explanatory power (how well it does account for a phenomenon) which has no cognitive dimension and for which veridicality is most often considered essential. Here veridicality refers to what I have called V1-veridicality. A V1-veridical explanation is one in which premises are true or well-confirmed and stand in the right kind of relationship to an explanandum such that they account for the explanandum.

In contrast, explanations that are offered in support of expert claims and AI outputs are explanations-as-communications. They are post hoc and partial accounts of reasoning and computational processes for which the cognitive dimensions of understandability and context sensitivity are key to explanation quality and for which explanation quality refers primarily to justificatory value (the extent to which the information conveyed by an explanation can help a recipient justify her (dis)belief in a claim), not explanatory power. These kinds of explanations can be veridical in two senses. V1-veridicality still pertains to how well an explanation does account for a claim – for example, whether the physician's explanation does account for the expected efficacy of an antibiotic treatment – though it is not essential to an explanation's justificatory value. V2-veridicality,

---

In a slightly different vein, Henk de Regt and Victor Gijsbers (2016) argue that one can remain agnostic about the influence of idealization on the explanatory power of a model while commenting separately on the influence of idealizations on the efficacy of the model with respect to other purposes they might serve. For example, they propose that scientific models can succeed in performing core tasks of science (pragmatic ends) such as making accurate predictions, guiding practical applications, or developing better science, irrespective of the satisfaction of any veridicality condition. They ultimately argue that these ends can be achieved by explanatory models "that are false, and not just slightly false, but wildly so," and present phlogiston theories of chemical phenomenon, Newtonian gravitational theory, and substance models of energy and electricity as illustrative examples (50).

on the other hand, is concerned with fictionalization and describes how faithfully an explanation recounts the original reasoning or computational processes to which it refers.

None of this is to say that philosophy of explanation has no place in a discussion about the justificatory value of AI and expert explanations. Rather, when explanations are discussed as post hoc and partial accounts of reasoning it is important to recognize that we are interested in a particular case that strays from the paradigm model. In this section I have shown that philosophical discussion about AI explainability needs to be carefully situated within the pragmatic branches of philosophy of explanation literature. AI explanations are means to an end (e.g. appraising system safety, looking for sources of bias, justifying belief in outputs, etc.) and a high quality explanation will be one that provides a recipient with information she can use to achieve that end.[9] This dissertation is concerned specifically with the pragmatic end of justified belief in AI outputs.

In the following section I investigate how abstractions in explanation reflectivity (V2-veridicality) and detail impact the justificatory value of explanations. To help guide the discussion, I reframe the justificatory value challenge within the literature about contexts of discovery and justification in scientific research. This discovery-justification framing does not replace or oppose relevant discussions within philosophy of explanation literature. Rather, the framing helps situate AI-to-user and expert-to-novice explanations within philosophy of explanation literature, keeping us engaged with the testimonial stance and therefore focused on the specific challenges and considerations relevant to the post hoc partial explanations we have in mind.

3. DISCOVERY, JUSTIFICATION & THE DOING-SAYING MISMATCH

In Section 1, I described how most XAI researchers adopt a kind of testimonial stance towards AI systems in that AI explanations are thought of much like human expert explanations. To begin, it is acknowledged that simplifying the AI system itself is not a viable option for achieving system explainability. This observation is in line with the unspoken assumption that where human expert reasoning processes are too complicated or high-level for a lay recipient to understand, the proper response is not to ask the expert to employ a simpler reasoning process (unless, of course, there is reason to believe a simpler course of reasoning would also be epistemically superior) but to find a more cognitively accessible way to account for her claim. Accordingly, AI explanations resemble human explanations in that both are post hoc – they do not provide a direct look at the expert/AI's internal reasoning/computational processes – and partial – they are in some way simplified or otherwise abstracted from the original reasoning/computation process for the sake of recipient understanding. The post hoc and partial nature of AI and expert explanations results in a doing-saying mismatch between what the AI/expert does to derive a conclusion and what she reports about that

[9] Rune Nyrup and Diana Robinson (2022) arrive at a very similar conclusion in their investigation of using philosophy of explanation literature to inform discussion about AI explainability in a healthcare context.

process. Determining justificatory value of an explanation is a question of navigating this doing-saying mismatch, and in this section, I describe how the debate about contexts of discovery and justification in scientific research can help frame the discussion.

The discovery-justification literature is, at its core, all about post hoc and partial explanations and about belief in scientific testimony. As Jutta Schickore (2008) explains, the discovery/justification (DJ) distinction was initially outlined by Hans Reichenbach (1938) to lay a framework for investigating the epistemic significance of the inherent mismatch between what scientists do in the lab to produce their findings and what scientists say about those processes in their academic and public communications. In the "context of discovery" scientists conduct experiments, analyze data, and formulate theories, and in the "context of justification" scientists communicate their findings beyond their immediate research group.[10] However, when scientists present their findings in papers and talks, the procedures they describe rarely accurately reflect the events and thought processes that led to the result obtained. Rather, the rationale and methodologies a scientist reports in the context of justification are, as Schickore describes, "post hoc rational reconstructions" of the scientist's process of discovery. Scientists change the order of their investigative steps, rationalize their choice of subject or experimental procedure, and omit details they find irrelevant all with an eye to supporting their findings in a way that is easily followed and robust to peer criticism. Accordingly, the question arises: *Is what a scientist says in a paper justificatorally valuable because of or despite any abstraction (in detail or reflectivity) from what she does to derive a conclusion?* Addressing this question will contribute not only to a foundational question of the discovery/justification literature, but also shed light on the parallel question posed by post hoc and partial explanations offered by AI systems. *Is what an AI system "says" justificatorally valuable because of or despite any abstraction (in detail or reflectivity) from what the AI system "does" to derive a conclusion?* There are two sides to the debate.

First is the negative view: explanations are only justificatorally valuable *despite* any abstraction from original reasoning processes. This viewpoint is most famously articulated by Peter Medawar. In his landmark BBC talk, *Is the Scientific Paper a Fraud?,* Medawar (1963: 33) states his view unapologetically; what scientists actually do in the lab provides the best justification for their outputs, while any changes a scientist makes to the report of their reasoning and procedures only degrades the justificatory value of their account. Medawar writes, "the scientific paper embodies a

[10] There is much disagreement among philosophers of science about whether the context of discovery or justification is a proper target for philosophical analysis. Schickore largely attributes this disagreement to the oversimplified interpretation of Reichenbach's DJ distinction commonly adopted by philosophers of science; most contemporary philosophers overlook Reichenbach's nuanced account of the contexts of discovery and justification bridged by rational reconstruction and credit Reichenbach only with outlining a binary distinction between the context of discovery (the empirical analysis of which is the role of psychologists, anthropologists, and sociologists) and the context of justification (the normative analysis of which is the role of philosophers). It is a dichotomy that is critiqued strongly (Kordig, 1978). However, Schickore points out that the motivation behind Reichenbach's initial work is unquestioned, scientists do not preach what they practice, and the DJ debate need not be resolved to investigate further the epistemic significance of the mismatch between what scientists do and what scientists say.

totally mistaken conception, even a travesty, of the nature of scientific thought," and therefore an epistemic evaluation of a scientist's findings ought to focus on the processes by which those findings were derived.

I suggest that those who take transparency as the ideal of AI explainability sit with one foot stuck in the design stance and implicitly think along the same lines as Medawar. They wish solely to appraise AI system functionality in terms of algorithmic mechanism and, as such, favor a direct view of those mechanisms unsullied by abstractions in detail and reflectivity. For example, in his differentiation between "how actually" and "how possibly" AI explanations, Christopher Grimsley (2021: 7) writes, "in the case of explanations of high-stakes automated decisions, 'how actually' should be the standard." Grimsley's concern is that fictionalizations "conceal the truth" behind AI outputs and can provide "the appearance of an adequate explanation with none of the substance". To provide another example, the UK House of Lords Select Committee on Artificial Intelligence (House of Lords, 2018: 39-40) describes explainable AI as "AI systems that are developed in such a way that they can explain the information and logic used to arrive at their decisions," and they argue that "where it is not yet possible to generate thorough explanations for the decisions that are made, this may mean delaying their deployment for particular uses until alternative solutions are found". As is standard in the XAI literature, the notion of explanation adopted by the House of Lords Select Committee refers to the internal processes by which a system reasons to a conclusion. While it is acknowledged that "full technical transparency" would be difficult if not impossible with regard to certain kinds of AI systems, especially for high stakes AI applications as in medicine or law, transparency remains an ideal to strive for. That transparency is taken as an aspirational gold standard implies that it is assumed that the justificatory value of an explanation is rooted in its fidelity to the computational process employed to derive a conclusion, and that abstraction from that original process detracts from the explanation's justificatory value.

On the other hand, philosophers such as Frederick Suppe (1998), Franklin and Howson (1998), and Peter Lipton (1998) defend the more optimistic position that post hoc reconstructions of a scientist's reasoning and procedures provides the best justification for the her findings; a scientist presents her most well-reasoned arguments in her published works, and it would be foolish to ignore these mindfully constructed justifications and attend only to a verbatim report of the scientist's original reasoning process. A published work contains additional information that can usefully inform one's appraisal of a finding – like references to supporting literature – while irrelevant and distracting information is omitted – like a daydream that inspired a hypothesis or the myriad of mistakes made in the process of establishing a new experimental protocol.[11] Adopting this viewpoint we might be more optimistic that explanations that do not provide a direct look into the processes by which a conclusion

[11] For example, August Kekulé's discovery of the molecular ring structure of Benzene was prompted by his dream of a snake looping around to bite its own tail (Rocke, 2010). This nice story about the early stages of Kekulé's reasoning process contributes nothing to substantiate his discovery.

was derived (either by a human expert or an AI system) can be justificatorally valuable because of the modifications that distance it from the original 'doing' process.

In addition to containing the most well-reasoned arguments for a conclusion, abstraction from original reasoning is key to an explanation's justificatory value in cases where it is not possible to provide a full account of the reasoning that led to a conclusion. Such is the case, for example, in massively collaborative scientific research projects in which a published paper is based on the work of numerous practitioners separated by significant temporal and geographical distance (Hardwig, 1991; Galison, 2003; Kukla, 2012; Winsberg, et al., 2014). No one author is privy to the full discovery process nor is it possible for all the decisions made and actions taken by each practitioner to be recounted in full. Rather, a post hoc justification for the finding and a report of the general research methods employed at various stages must suffice.

The same is also true of many cases of individual reasoning. As John Zerilli et al. (2018) explain, especially in areas where intuition is a crucial element of expert decision-making, the expert relies on subdoxastic factors – reasons that exist below the level of consciousness – to guide her reasoning and therefore will not be fully aware of, nor able to rearticulate, the reasoning she employed (see also Dreyfus & Dreyfus, 1986).

In what follows I begin to investigate the justificatory value of expert-to-novice and AI-to-user explanations in accordance with the optimist's perspective. I ignore Medawar's pessimistic view if only because it ends the discussion too quickly. If an AI explanation or a scientist's paper only retains its justificatory value despite its abstractions from the original process by which an output or finding was derived, then AI explanations are necessarily of low justificatory value. This is because the more complex an AI system (or the more complex an expert's reasoning process), the more abstractions in detail and/or reflectivity that will be necessary to make an explanation cognitively accessible. As such, the only way to increase an AI explanation's justificatory value is to reduce the complexity of the system thus making its internal processes more easily explained without abstractions. This is, however, an undesirable solution due to the loss in AI system power that accompanies reductions in complexity. Under the optimist's perspective, however, it is still possible that an AI system's complexity (or the complexity of expert reasoning) is not an immediate death warrant to the justificatory value of an explanation-as-communication. If there is a potential for abstractions to contribute to justificatory value, then there is still room for more complex and powerful AI systems to be able to produce explanations of high justificatory value. If, however, under the optimist's perspective it still comes to light that there are cases in which explanations are of minimal justificatory value, then we are necessarily in serious trouble in terms of our ability to justify belief in either the outputs of AI systems or the claims of human experts.

I now discuss how fictionalization and idealization can influence the justificatory value of the post hoc and partial explanations offered by AI systems and human experts. Each can contribute to or

detract from the justificatory value of an explanation in different ways which are summarized in Figure 1.2 near the end of this section. I begin with idealization.

Idealization (abstraction in detail)

The more detailed an explanation is the more information it contains for an explainee to appraise. An explanation's justificatory value refers to the degree to which it conveys information that can help an explanation recipient epistemically justify her (dis)belief in the claim the explanation supports. Therefore, the more detailed an explanation the higher the potential justificatory value of the explanation will be. This is, of course, assuming that those details are in some way relevant to the epistemic well-foundedness of a claim, for instance, excluding inspiring daydreams or other tangentially related facts that have no bearing on the veracity of the claim.

Note, however, that there is a limit to the justificatory value of an explanation as constrained by the cognitive capacities of the explanation recipient. I call this an explanation's *maximum potential justificatory value*. If the recipient has a limited capacity for understanding as restricted by the relevant background knowledge she holds and by her cognitive capacities, then she will only be able to do so much to use the information provided to her in an explanation to mount her own appraisal of the claim and decide whether to believe it.[12] For example, suppose an AI system were to present a list of the statistical weights assigned to its input factors by way of an explanation. Some people, perhaps a statistician, would be able to use that information to make an appraisal of the output, but for others the explanation is too complex and/or beyond the kind of evaluation their background knowledge

[12] As I speak about capacity for understanding, this is a good place to make a nod to the extensive epistemological literature on understanding in science, its nature, and sources. I recommend de Regt, Leonelli & Eigner's (2009) edit collection, *Scientific Understanding: Philosophical Perspectives,* for a thorough overview on the subject.

In this dissertation I largely steer clear of the literature because it is primarily concerned with (and enjoys very little agreement on) how understanding is defined, how understanding can be conveyed, and if and how it is distinct from knowledge – that is, whether one can have understanding without knowledge and whether understanding is epistemically inferior or superior to knowledge (Khalifa, 2013, 2017; Pritchard, 2008; Kvanvig, 2003; Lipton, 2009; Grimm, 2006, 2016). I need not engage deeply with these discussions for the purpose of my own investigation into the justificatory value of explanations. I will, however, briefly describe a differentiation between two senses of understanding that is consistent throughout the literature, and I will note which one I attend to.

The delineation is between the phenomenological *feeling* of understanding (i.e. the "aha!" experience that accompanies the moment when someone suddenly "gets" what is being explained to them) and the idea of understanding as something more than a mere feeling (de Regt, 2009a, 2009b; Trout, 2002, 2007; Grimm, 2010). Thoughts on what that 'something more' is, vary, but they often have something to do with one's "grasp" of or ability to produce a true explanation for a phenomenon and/or one's ability to use the information she supposedly understands to achieve some other end. For example, if a person understands an explanation for an event, she would be able to see how things could have turned out differently or why they had to happen the way they did.

I think about understanding in the second sense as something more than a mere feeling. I assume that if a person understands an explanation (here, a post hoc account of reasoning) she is able to make use of the information it contains to help justify her belief. Note, however, that I do not connect understanding with knowledge. I allow that a person's understanding can be false (i.e. a misunderstanding) in which case the explanation will fail to have justificatory value for that recipient. I expand on this point in the following pages.

would enable. In this way, recipient capacity for understanding places an upper limit on an explanation's justificatory value. In other words, the same explanation can hold high justificatory value when offered to an expert audience (like a scientific paper shared by the author to another leading expert in the field) and low justificatory value when offered to a non-expert audience (like a scientific paper shared with an undergraduate student). The idea that recipient capacity for understanding caps the potential justificatory value of an explanation plays a central role in the following chapter. Now I turn to the influence of fictionalizations (abstractions in explanation reflectivity) on the justificatory value of explanations.

Fictionalization (abstraction in reflectivity)

As a reminder, an explanation is less reflective the more it distorts an account of an original reasoning process, for instance, by introducing fictionalizations like analogies, examples, or other thought processes that were not used by the explainer to derive her conclusions. Medawar's complaint against the fraudulence of scientific papers, for example, is primarily a complaint about fictionalizations – about how little scientific papers reflect the processes by which scientific findings are actually derived. Not only do the papers leave out details about the original discovery processes, but the post hoc rationalizations they include fictionalize scientific papers in the sense that the original discovery process cannot be reconstructed by simply adding back in omitted content.

In this way abstractions in reflectivity are independent from abstractions in detail. It is possible, for instance, for an explanation to be highly reflective but low in detail. For example, an explanation for how an AI system derived a specific diagnosis that reads, "the diagnosis is the best fit for the inputted data," does not distort the system's reasoning process, but it also contains very little information about the process by which the diagnosis was produced that an explanation recipient could use to evaluate the claim. On the other hand, explanations can also be high in detail but low in reflectivity such as when an expert provides post hoc rationalizations for decisions initially based on intuition. For example, an experienced chiropractor might intuitively feel how to place her hands and apply pressure to adjust a patient's spine. If asked how she decides to perform adjustments, she could say "I find what feels right" (low detail, high reflectivity) – which may be an accurate representation of the chiropractor's decision-process in the moment but is not very helpful to an explainee trying to decide whether to accept the chiropractor's treatment. The chiropractor is more likely to reference the anatomy of the musculoskeletal and nervous systems and describe how specific pains can be alleviated by relieving pressure between joints by performing certain manipulations (high in detail, low in reflectivity). In so doing, the chiropractor also provides the patient with more information the patient can use to help epistemically justify her acceptance of the chiropractor's treatment recommendation than if the chiropractor had just referenced her intuition.

In addition to providing additional information for explanation recipients to evaluate when the original reasoning process is lacking, less reflective explanations can also be used to present a reasoning process in a way that is better understood by the audience where a more highly reflective account would have to be severely simplified and therefore capped at a lower justificatory value. This is particularly the case for AI explanations. AI systems are not designed to reason like humans but to mimic the outputs of human expert thought by statistically 'crunching' vast sets of data. Accordingly, AI generated explanations that are highly reflective of the internal processes employed by the system to derive the output do not appeal to human norms of reasoning. As Tim Miller (2018) explains, humans are naturally predisposed to reason causally, by example, and by analogy and therefore explanations that are most easily understood and appraised by humans are not highly reflective of the AI system's reasoning processes. Examples include image classification systems that explain decisions in terms of shape and color rather than pixel proximity, or cancer treatment planning systems that offer examples of similar prototypical diagnostic cases instead of quoting statistical analyses. Miller argues that any conception of good explanations offered by XAI researchers should directly respond to literature in psychology and cognitive science investigating how people are naturally inclined to generate, select, evaluate, and present explanations to one another. In the same vein, David Leslie (2019: 47) writes:

> This cognitive dimension has a direct bearing on how you should think about offering suitable explanations about algorithmically generated outcomes: Explaining an algorithmic model's decision or behavior should involve... making intelligible to affected individuals the rationale behind that decision or behavior *as if* it had been produced by a reasoning, evidence-using, and inference-making *person.* (Emphasis added)

By altering reasoning accounts to better align with norms of human reasoning, more detailed explanation can be provided to explanation recipients where more reflective explanations would have to be significantly simplified and therefore capped at a lower justificatory value. For example, a prototype explanation is a kind of example-based AI explainability strategy (also called case-based explanation) in which particular instances of a data set that are considered "prototypical" of the data set are selected to explain the AI system's output. The idea is that a new data point's membership to a category can be explained by its similarity to the prototype data point (Adidi & Berrada 2018; Guidotti et al., 2018; Bein & Tibshirani, 2011). For example, a medical diagnostic system might present two or three prototypical patient case files to a physician to explain the diagnosis (e.g. a classification of cancerous or non-cancerous) it outputted for a new patient. The physician can appraise the similarities and dissimilarities between her patient's symptoms, medical histories, test results, etc. and those of the prototype cases to better understand why the AI system outputted the diagnosis it did and, in turn, to evaluate by comparison to her own medical knowledge and experience whether she ought to believe the AI system's output.

To be clear, the AI system does not derive its conclusion by comparing the new patient data to the two or three prototype cases selected. The prototypes are only meant to be representative of cases

in the data set that have been classified a particular way. Rather, the AI system's internal processes are "fiendishly intricate" statistical analyses (to use Zerilli et. al.'s (2018) term) in which vast set of data are passed through complex layered networks of weighted nodes and edges in order to determine the statistically optimal classification for the new data points pertaining to the patient in question (Zerilli et al., 2018; Burrell, 2016; Price, 2015; London, 2019). If the physician could track and comprehend the whole statistical process, that might give her strong reason to believe (or disbelieve) the output, but the most simplified version of that process that the physician can understand is most likely something along the lines of, "the system was presented with thousands of data points from past patients which it ran through a complex iterative statistical analysis to output diagnosis X". While reflective of the AI system's reasoning process, such an explanation provides little information a physician could use to evaluate for herself – based on her own medical knowledge and experience – how likely the output is to be true. A couple of concrete prototypical examples drawn from the data, however, are more understandable to the physician and more easily compared to her own knowledge and experience. In this way, a less reflective explanation can be more justificatorally valuable than a more reflective alternative.

That all being said, greater reflectivity also has its advantages with respect to the justificatory value of an explanation. To begin, reflectivity of original reasoning is important if the explanation recipient is trying to determine if and how she ought to intervene on that reasoning process. For example, AI safety researchers investigate potential sources of bias and error in AI reasoning, and peer reviewers of scientific papers look to identify any questionable research practices and to request changes to protocol or experimental design to ensure that proffered results are robust. Overall, the concern is that in prodding at completely fabricated accounts one might identify errors or biases in reasoning that do not exist or they might propose solutions or revisions that do not yield real improvements because they do not address the root of the problem. Consider, for example, the case in which a researcher sets out to reproduce a scientific study. If the published study does not recount the original experimental procedures both with sufficient detail and reflectivity, the researcher may find the procedures difficult to reproduce and, consequently, the findings difficult to either confirm or disconfirm.[13]

Furthermore, only reflective accounts of reasoning can provide information that might be used to evaluate the safety and sensitivity of initial reasoning processes. Safety and sensitivity are concepts typically discussed in the context of what is now called anti-luck epistemology (Pritchard, 2007), a field that builds on Edmund Gettier's (1963) observation that the Justified True Belief (JTB) concept of knowledge frays where justified true belief can be acquired by luck. For example, following on Gettier, Alvin Goldman (1976) explores a thought experiment in which a person believes there is a barn in a field because she sees a barn, but what she in fact sees is a perfect barn façade. Yet,

[13] For more on the replication crisis in scientific research, its existence, causes, and epistemic implications *inter alia* see Ioannidis (2005), Baker (2016), Radder (2003, 2006), Schmidt (2009), and Fidler & Wilcox (2021).

her belief that there is a barn in the field is still true because, by a stroke of luck, there happens to be a very real barn hidden behind the façade. Her belief that there is a barn in the field is true, and, one might argue, justified (there is no reason not to believe the perfect façade barn is indeed a barn), but does it constitute knowledge?

Sensitivity theorists would answer that the person does not know there is a barn in the field because the process by which she formed her belief is not sensitive to the world. That is, in possible worlds where her belief is false (worlds in which people build perfect barn façades to trick passersby), she would still come to hold the same belief (that there is a barn in the field). As originally articulated by Robert Nozick (1981) a person's belief that *p* is sensitive only if, if *p* were false, the person would not believe that *p*. John Greco (2012: 194) nicely captures the essence of the sensitivity condition of knowledge in the following statement: "The spirit of a sensitivity condition is that, in cases of knowledge, one would notice if things were different…. More generally, when one is sensitive to some state of affairs, one would react to a difference".

Another response is to posit that beliefs must also be, or instead be, safe to constitute knowledge. As described by Timothy Williamson (2000) a true belief is safe and therefore constitutes knowledge only if in nearby worlds where a person believes that *p*, *p* is true. The safety condition of justification captures the pragmatic consideration that the closest possible world in which a person's belief is false might be quite distant. A safety theorist might respond that your belief that there is a barn in the field does constitute knowledge because a world in which perfect barn facades are constructed to deceive passersby is a distant and unlikely world. Believing that what one perceives to be a barn is actually a barn is generally a safe bet. On the other hand, if a person believes that she has bought a winning lottery ticket, even if she later finds that she is correct, her belief is not knowledge because the worlds in which she is wrong are numerous and very close. Again, Greco nicely captures the essence of the safety condition: "The spirit of the safety condition is that, in cases of knowledge, [a person] would not easily go wrong by believing as she does" (2012: 194).

There is much debate between defenders of the sensitivity condition (Nozick 1981; Becker 2009; Black & Murphey 2007; Black 2008; Roush 2005) and those of the safety condition (Sosa 1999a, 1999b; Pritchard 2005b; Williamson 2000; Greco 2012) of knowledge. Some argue that a belief is safe because it is sensitive while others argue that if we pragmatically restrict the worlds we consider to those that are near, sensitivity might be recategorized as a subcategory of safety (Greco, 2012). I will not settle any debates here about the superiority or preferability of one condition over the other. Nor, as prefaced in the introductory chapter, am I interested in the precise conditions necessary for knowledge to obtain. Therefore, I am not interested in what particular degree or kind of sensitivity or safety is required for a person's belief to constitute knowledge. Rather I borrow the concepts of safety and sensitivity as useful tools for understanding how explanations can help a recipient acquire any amount of epistemic justification for her belief in a claim or output.

The idea is that the better the evidence an explanation provides to an explainee indicating that an output is sensitive and/or safe (or insensitive and/or unsafe), the more justified the explainee is in believing (or disbelieving) the output. For example, imagine that Becca claims that there is a barn in the field. She explains her belief by saying, "Via repeat observation I have discovered that the shape of shadow cast by a real barn (wide shadow) or barn façade (narrow shadow) in the morning or evening is a reliable indicator of which it is. It is now 7pm and that barn is casting a wide shadow, so I believe it is a real barn." Assuming that shadow shape is indeed a reliable indicator of whether a barn is real or a façade, Becca's explanation indicates that her belief is sensitive because it is based on relevant information (shadow shape and time of day) and safe because she is unlikely to go wrong in reasoning about barns in the way she does. As such, Becca's explanation provides us with good reason to believe that her claim is likely true. If, however, we have reason to believe the Becca's foundational premise about shadow shape is flawed – for instance, we know a barn façade will only cast a narrow shadow if it faces North or South but will cast a wide shadow if facing East or West – then we have reason to believe that Becca's belief is not sensitive. It fails to consider key, difference-making information. Furthermore, if we are aware that the barns in this area tend to be oriented every which way, then we know Becca's belief is unsafe. She could very easily be wrong.

Now to bring the discussion back to explanation reflectivity. If a post hoc account of reasoning does not closely reflect the original reasoning process employed by an expert or AI system to derive a conclusion, that explanation's ability to provide insight to the safety and sensitivity of the reasoning/derivation process in question is hindered. For example, if a human physician provides a minimally reflective explanation to account for a medical diagnosis, the recipient may struggle to evaluate, based on that explanation, whether the physician's reasoning processes is sensitive – i.e. whether the physician factored true and relevant information into her reasoning process about the patient's symptoms, test results, and any pre-existing conditions – and whether those reasoning processes are safe – i.e. whether, given the relevant inputs, the physician's reasoning process is one that is likely to produce correct diagnoses. Similarly, in the case of the kinds of AI prototype explanations described above, by being provided with two or three patient case studies ("prototype data points") a physician may get a rough sense that the AI system factored relevant information into its diagnosis (an indication of system sensitivity), but the physician receives no information about how that data was processed to derive the conclusion. As such, she cannot make any determinations about the safety of the AI system's diagnosis derivation processes – that is, whether the derivation process is one that is likely to produce correct diagnoses given relevant input data such that it would be hard for the user to go wrong in believing the system's outputs. She might be able to draw a conclusion about the system's safety based on its track record and past performance, but to glean this kind of information from an explanation she would need to be provided with, and be able to comprehend and appraise, a more reflective and sufficiently detailed account of how the AI system did derive its conclusion. The challenge, of course, is that any highly reflective explanation for an AI

output that also contains sufficient detail to shed light on the reliability of the AI's derivation process will likely be too complex for the typical user to comprehend.

To summarize, the justificatory value of an explanation will vary with alterations to the explanation's detail and reflectivity. Reductions in explanation detail are prerequisite to justificatory value where a simplified account is needed to achieve cognitive accessibility. However, for those who would be able to comprehend fuller accounts, reduction in detail places an upper limit on the potential justificatory value because less information is contained in the explanation for the recipient to analyze. In a similar trade off, more reflective explanations allow recipients to evaluate the safety and/or sensitivity of the explainer's reasoning process. However, reductions in reflectivity can also boost justificatory value by presenting information in a way better understood by the audience where a more reflective account would have to be severely simplified for the sake of recipient comprehension, and therefore capped at a lower potential justificatory value.

[Figure 1.2] The influence of explanation detail and reflectivity on justificatory value

	Contributions to justificatory value	**Detractions from justificatory value**
Detail	▪ More information to analyze allows for higher justificatory value.	▪ Increased complexity can make it more difficult to understand. Recipient capacity for understanding may therefore cap the maximum potential justificatory value of an explanation.
Reflectivity	▪ May provide indication that the original reasoning process is safe and sensitive. ▪ Allows the original reasoning process to be prodded for sources of error or bias.	▪ Less reflective explanations may frame an explanation in a way more easily understood. ▪ More detailed explanations capped at a higher justificatory value may be available that do not reflect original reasoning processes.

I conclude this section by noting an implication of the section's discussion for the supposed AI explainability-performance trade-off. Because it is possible for abstractions in detail and reflectivity to contribute to the justificatory value of an explanation, it follows that there may also be cases in which the trade-off dissolves. If an explanation can be justificatorally valuable without being highly reflective of the internal process employed by the AI system, then there is no need to simplify the AI system itself in order to make it explainable. We may focus instead on finding other ways of

explaining an AI system's derivation process that best help the user epistemically justify her (dis)belief in the system's outputs given the limitations on her capacity to comprehend.

However, while there is not necessarily a tension between an explanation being able to have high justificatory value and the complexity (and correspondingly the performance) of an AI system, I caution that we should not assume that there is never such a tension. First, recall that there are instances in which reflectivity can be epistemically useful, for instance, by illustrating that the explainer used relevant evidence to inform her reasoning process thus showing that her reasoning process is sensitive., it should also be emphasized that there are cases in which reflectivity plays important non-epistemic roles. Put another way, even when a more reflective explanation would not be of any more help in acquiring epistemic justification for belief, there still may be good reason to prefer an explanation that is faithful to an original reasoning or derivation process and hence, an argument for using simpler AI system for which detailed and reflective yet comprehensible explanations can be offered.

One non-epistemic benefit of high reflectivity is that it enables appraisal of the fairness of decision-making procedures. This is a key concern, for instance, when using AI systems to make hiring or school admissions decisions (Köchling & Wehner, 2020; Ötting & Maier, 2018). For example, imagine two expert systems (systems A and B) are used to make a hiring decision for two different companies. Both systems reject applicant X who has applied for a job at both companies with the same CV. Both systems also accept applicant Y. (For argument's sake, say both systems always come to the same decision). System A bases its analysis solely on information in the applicants' CVs regarding grade point average (GPA), level of education, and years of prior employment. Using system A, applicant X made it to the final round of cuts where she was eliminated based on her GPA. System B analyses the same metrics as system B in addition to scanning candidates' CVs for information about extracurricular activities. Due to biases in the system's training, system B first cuts all candidates who list involvement in a women's sports team. Using system B, applicant X was cut in this early round. While systems A and B always make the same decision, there is an argument to be made that in case B, but not case A, applicant X was treated unfairly. Gender is a legally protected category, yet in case B the candidate was indirectly penalized for being female. Only through the analysis of explanations that reflect the influence of "women's sports" on CV selection could the source of unfair treatment be identified and corrected.

The point here is that the importance of reflectivity depends on the purpose for which the explanation is being used. If, for instance, an explainee is interested in the procedural fairness of an AI system's decision-making process, then a higher degree of reflectivity may be called for. But where the goal of an explanation is to provide people with information they can use to help epistemically justify their belief in a claim, a highly reflective explanation can, but will not necessarily, present a recipient with information she can understand and evaluate. Especially where an explanation

recipient's capacities for comprehension and analysis are limited, a less reflective explanation may be more well suited to the task.

Now having outlined the different impacts abstractions in detail and reflectivity can have on the justificatory value of an explanation offered in support of a claim or output, in the following chapter I answer the looming question: just how justificatorally valuable can we expect AI and expert explanations to be for lay users and novice recipients?

4. CONCLUSION

The overarching goal of this dissertation is to determine how people can acquire some epistemic justification for their (dis)belief in AI outputs. Given the breadth of applications to which AI-enabled expert systems are being employed, being able to justify belief in this way is an increasingly large part of maintaining a society's epistemic security – that is, the ability of people in the society to consistently distinguish between true information and false or misleading information.

This chapter progressed in three main steps. First, in Section 1, I presented and defended the idea of adopting a testimonial stance towards AI systems and other information producing and mediating technologies. In contrast to Dennett's design stance in which instruments are appraised merely in terms of their design and functionality, the testimonial stance involves thinking about how one might justify her belief in AI system outputs by comparing to how people justify their belief in human speaker claims.

In this and the following chapter I am specifically interested in how the testimonial stance helps us think about the role of explanations in helping users to justify their beliefs in AI outputs. Overall, adopting the testimonial stance means viewing an AI explanation as a post hoc and partial account of reasoning like an explanation offered by a human expert to a novice. These explanations are not direct windows to viewing how the system works, like lifting the hood of the car, rather they are offered after the fact like a mechanic's report of the state of a car engine or a scientist's publication of her discoveries and research methodology. As such, there is a gap between the original output derivation process and the provision of an explanation thereof in which some decisions must be made on the explanation recipient's behalf about what information she is to receive and what information will be simplified or otherwise altered for the sake of understandability. This results in a doing-saying mismatch between what an expert/AI does to derive an output and what she/it 'says' about that process to an explanation recipient. The doing-saying mismatch in turn raises questions about how alterations to an original reasoning process affects the explanation's justificatory value – its ability to provide the recipient with information she can use to help epistemically justify her (dis)belief in the output in such a way that the explanation helps point the recipient in the right direction – that is, toward believing that which is true and away from that which is false.

In section 2 I turned to literature on the philosophy of scientific explanation to help elucidate the features of justificatorally valuable explanations. I cautioned, however, that the notion of explanation as understood by philosophers of science – a non-cognitivist notion of explanations-as-accounts that must meet some veridicality condition for explanatory correctness – does not align with our idea of explanations a post hoc and partial accounts of reasoning. This is not to say that philosophy of explanation has nothing to offer, rather that we need to be careful to situate our conversation according to cognitivist notions of explanation-as-communications and to keep in mind the specific pragmatic end we are hoping explanations will achieve – providing justification for belief. Overall, we must remember that a justificatorally valuable explanation will be one that presents an account of reasoning that is comprehensible to a recipient and which contains information the recipient can use to evaluate the veracity or well-foundedness of a claim using the cognitive capacities and background knowledge available to her.

Finally in section 3 I proposed we turn to a similar discussion taking place in the discovery-justification literature to help guide an investigation of how abstractions for original reasoning/computational processes affects the justificatory value of an explanation. Debates around contexts of discovery and justification in scientific research are not often linked to issues of explanation and testimony, but the driving motivation behind the literature is navigating the doing-saying mismatch inherent to scientific publications, expert communications, and AI explanations. I analyzed two factors, abstractions in detail and abstractions in reflectivity (analogous to the concepts of idealization and fictionalization as borrowed from the literature on scientific models), for how they might both positively and negatively contribute to justificatory value. I note that the cognitive capacities and resources of the explanation recipient may also place an upper limit on the justificatory value of an explanation. The next chapter gets to the meat of the issue: what *can* we expect the justificatory value of AI-to-user and expert-to-novice explanations to be?

2

Continuums of Justificatory Value

The aim of Part 1 of this dissertation is to investigate the justificatory value of AI and expert explanations. I consider an explanation to be justificatorally valuable insofar as it conveys information that, via (in)coherence with an explanation recipient's background knowledge and beliefs, helps the recipient to acquire some degree of internal epistemic justification for her (dis)belief in the explainer's claims, conclusions, or recommendations. The idea is that if explanations can provide users/novices with information they can use to appraise for themselves whether an AI output or expert claim is likely to be true, then explanations can be used to facilitate more epistemically secure AI-to-user and expert-to-novice communications.

In the previous chapter, I outlined my position on the debate as to whether explanations (understood as post hoc and partial accounts of reasoning) are justificatorally valuable because of or despite any abstractions from the original reasoning processes by which the claim or output was derived. In short, the cognitive accessibility of an explanation is prerequisite to justificatory value. As such, I argue (1) that reductions in detail are often necessary for an explanation to be understandable to a recipient, however (2) reductions in detail below that which is comprehensible to a recipient decrease the justificatory value. (3) Abstractions in the explanation's reflectivity (its fidelity to the original reasoning process) can also detract from justificatory value because less reflective explanations cannot provide as much evidence of the safety or sensitivity of a reasoning process as a more reflective explanation would. However, (4) it is also possible for abstractions in reflectivity to contribute to justificatory value by presenting information in a way better understood by the audience

where a more reflective account would have to be severely simplified and therefore capped at a lower potential justificatory value in order to cater to recipient understanding.

With these influences on justificatory value in mind, I now set out to answer the question I originally posed: What can we expect the justificatory value of expert-to-novice and AI-to-user explanations to be? Or put another way, how effective can we expect expert-to-novice and AI-to-user explanations to be at helping recipients make epistemically well-informed decisions about whether to believe expert claims and AI outputs?

In Section 1, I first described how coherence with background knowledge is the root of an explanation's justificatory value. As part of this discussion, I explore how three AI explainability strategies might yield explanations of some justificatory value, and I address a key complaint against coherence theories of epistemic justification. In Section 2, I then discuss how the justificatory value of an explanation is limited by the recipient's cognitive capacities and background knowledge. Accordingly, I argue that the more epistemically imbalanced an explainer-explainee relationship is, the lower the maximum potential justificatory value of an explanation. This means that many explanations offered by experts to novices, or by AI systems to lay users, will necessarily be of low justificatory value. In Section 3, I comment on the justificatory value of completely fabricated explanations – these are explanations which in no way reflect the reasoning or computational process by which a claim or output was derived (a completely non-V2-veridical explanation). Section 4 concludes both this chapter and Part 1 of the dissertation.

1. JUSTIFICATORY VALUE OF EXPLANATIONS VIA COHERENCE

An explanation is justificatorally valuable insofar as it conveys information that, via (in)coherence with an explanation recipient's background knowledge and beliefs, helps the recipient to acquire some degree of internal epistemic justification for her (dis)belief in the explainer's claims or outputs. As discussed in the introductory chapter, I am interested in internal epistemic justification as opposed to external justification because this dissertation deals with the practical issues of informed decisions-making, specifically how one informs her decisions about whether to believe the claims and outputs of experts and AI systems. If a person engages in informed decision-making, it implies that she holds internal reasons for her decisions (to believe), and that her decisions (to believe) are not made on a whim. To be clear, when I refer to a person's internal epistemic justification for belief, I do not refer merely the subjective feelings of confidence one might have in her beliefs. A person might feel confident without holding any reason at all, and it would be a hard case to sell that the phenomenological experience of confidence is a good justification for belief in and of itself. Rather I understand internal epistemic justification in an objective sense as a measure of how well a new belief coheres with one's existing background knowledge and beliefs.

Briefly put, epistemic justification via coherence depends on how a belief supports, is supported by, or otherwise dovetails with existing beliefs in the set (Quine & Ullian, 1970; Thagard, 2005). More specifically the coherence of a set of beliefs is a function of logical consistency and explanatory connections between beliefs. A set of beliefs is logically consistent if none of its members are mutually exclusive, and explanatory connections have to do with how well the beliefs fit together and support one another. The more the addition of a new belief to a set increases the number and strength of explanatory connections in the set, the more justified the belief (Feldman, 2003; Lemos, 2012). For example, if Katy's digital weather monitor tells her that it is raining, and Katy (a) heard on the morning news that there is a 60% chance of rain in the afternoon and (b) Katy has a sharp pain in her elbow (as often happens when the weather changes), Katy is more justified in believing the monitor's announcement that it is raining than if she had only felt the pain in her elbow or she had only heard the morning's weather report.

There is, however, an obvious complaint against coherence with pre-existing beliefs as a basis for epistemic justification. What is often called the 'isolation objection' to coherentism notes that coherence is a condition internal to a set of beliefs. Therefore, there is little reason to think that a set of beliefs reflects reality simply because that isolated set is internally coherent.[1] Rather, if a person's pre-held beliefs are false or unjustified, then coherence with those beliefs is not an indication that those beliefs are epistemically well-founded or likely to be true.

Bear with me while I set this complaint aside for just a moment longer. I will return to it at the end of this section. I would first like to discuss what it looks like for coherence to be the root of the justificatory value of explanations so that I may also comment on the implications of the isolation objection for the justificatory value of explanations as well.

An explanation provides an explainee with information about the explainer's reasoning processes that the explainee can compare against her own beliefs, background knowledge, and understanding of what good reasoning looks like. The more relevant background knowledge she holds, the more capable she will be of identifying premises that are likely to be false – premises that do not cohere with the relevant knowledge she holds on the topic – or problematic steps in reasoning.[2] In this way, in addition to gaining some internal justification for believing a claim via that claim's

[1] For an extended discussion on the isolation objection to coherence see Haack (1993), BonJour (1985), Feldman (2003), and Lemos (2012). This objection is commonly posed by foundationalists who hold that the only justified beliefs are self-evident truths (also called non-inferential knowledge) or beliefs that rest on a foundation of self-evident truths in that they are either deduced or inferred from those truths (Newman, 2019). It is not necessarily the case that foundationalists deny that a first step toward justification can be coherence with one's background beliefs, but it is the foundationalist's view that we would need to ask whether those relevant background beliefs can in some way be reduced to self-evident truths which are externally justified. For more on coherentism vs. foundationalism debate see Sosa (1980).

[2] For example, Alvin Goldman (2001: 94) differentiates between *exoteric* and *esoteric* statements that form a full explanation or argument. Esoteric statements are those that are readily accessible to novices while exoteric statements pertain to the domain of expertise. The more esoteric statements an explainee is familiar with, the better placed she will be to appraise the explainer's reasoning process.

coherence with existing beliefs, justification can also be gained by appraising the reasoning process leading up to the claim's, or output's, derivation.

Consider, for example, how coherence of reasoning processes with existing beliefs plays a key role in peer review of scientific papers. A reviewing scientist holds background knowledge in a specific domain of expertise, experience in the discipline, and an understanding of proper scientific practice and methodology in her field. Even though a scientific paper is a post hoc reconstruction of the authors' original reasoning and investigative processes, it can still be evaluated for coherence with reviewer's existing beliefs about the knowledge that is already established in the field and about how experimental work is best conducted in the areas. For example, the reviewer can appraise whether the paper is grounded in the most relevant and well-established research in the field. Furthermore, given that a scientific paper supposedly presents the author's most well-reasoned justifications for her claims, if that post hoc reconstruction of reasoning does not cohere well with the reviewer's background beliefs about the field or about what good reasoning looks like in the area, then the reviewer would be less justified in believing the end claims. In other words, she would have good reason to reject the paper or to request revisions (e.g. changes to methodology or experimental set-up, additional experimental controls, further data collection, etc.). For example, the authors might describe causal interaction between cell signaling pathways that conflict with the reviewer's understanding of similar molecular mechanisms, or the authors might use experimental methods or analysis techniques that do not align with the reviewer's understanding of best practice in the field. For instance, perhaps the authors fail to describe both the positive and negative controls necessary to confirm that the observations they report are the result of the experimental conditions they propose and not some extraneous factor.

There is of course the possibility that an author simply fabricates her results. For example, a scientist might claim that result Y was yielded by using methodology X. If the reviewer believes methodology X to be an epistemically robust protocol, then even if the author made up her result, the reviewer's internal epistemic justification for believing Y would be boosted. Alternatively, an author might have employed questionable research practices which are difficult to spot such as data dredging (scouring data to identify statistically significant patterns to underpin retrospective hypotheses) or selective data reporting. Such practices are suspected contributors to the replication crisis in science (Ioannidis, 2005; Simmons et al., 2011; John et al., 2012). While it certainly is the case that some dishonest scientists will fall through the cracks, I posit that an experienced scientist (an expert reviewer) would still be able to sense more so than a layperson when results are being fabricated or when questionable methodologies are employed by picking up on small inconsistencies in reasoning or data that look too perfect or convenient. I return to the topic of completely fabricated explanations in Section 3.

Coherence with post hoc accounts of reason can play a similar role in underpinning epistemic justification for belief in AI outputs. This idea may seem strange at first glance. After all, as discussed

in Chapter 1, the challenge posed to XAI researchers is precisely that the internal computational mechanisms by which AI systems derive their outputs are often alien to natural modes of human reasoning, not to mention inaccessibly complex. As such, there appears to be a tension between the "otherness" of AI systems and the idea that explanations offered by AI systems and by human experts can be of similar use. Recall, however, that in Chapter 1 I have also shown that despite any differences in how AI systems and human experts reason, AI explanations are akin to human expert explanations in that both are post hoc and partial accounts of the original derivation/reasoning process employed by the AI/expert. Both human expert and AI explanations are simplified and modified to cater to recipient comprehension, and so, in both cases, the same question stands as to how abstractions in explanation detail and reflectivity effect an explanation's ability to provide recipients with information they can use to justify their belief in AI/expert outputs. More so, I have discussed how in both the human and AI case, there are times when reductions in explanation reflectivity (that is, how faithfully the explanation reflects the original reasoning/derivation processes) may be necessary for the explanation to hold justificatory value for the recipient. Such is the case, for instance, when experts arrive at conclusions via intuition. Accordingly, that an AI system reasons in ways unfamiliar to a user does not pose a hurdle to the idea that both AI and expert explanations can similarly be of justificatory value to users and novices. In what follows, I briefly describe three examples of AI explainability techniques that might be justificatorally valuable via coherence with recipient background belief even where the explanations include abstractions in detail and in reflectivity. These explanation types are as follows: prototype explanations, feature salience explanations, and counterfactual explanations.

I have already introduced prototype explanations in Chapter 1. In short, *prototype explanations* are a kind of example-based AI explainability strategy in which a data point that is considered "prototypical" of all data points that are classified in a certain way is provided as an explanation for why a new data point was granted or denied membership to the group (Bien, J. and Tibshirani, 2011; Kim et al., 2014; Gurumoorthy et al., 2017). For example, a medical diagnostic AI system might provide a physician with the case files of three past cancer patients who received a certain diagnosis to explain why a new cancer patient has received the same diagnosis. The idea is that the physician can appraise the similarities and dissimilarities between her patient's symptoms, medical histories, test results, etc. and those of the example cases to better understand why the AI system outputted the diagnosis it did. In turn, the physician can compare this information to her own medical knowledge and experience to decide whether to accept or reject the AI system's recommendation. To be clear, AI systems do not reason by example as humans often do. AI systems process vast sets of data via complex statistical analysis. The provision of prototype classifications describes a different reasoning method than that employed by the AI system, but it is a familiar form of reasoning that humans can more easily appraise.

Before I continue, a brief interjection; it should be acknowledged that a challenge arises when an explanation suggests to a recipient that an output ought not be believed, yet the system has a great track record for making accurate predictions, classifications, etc.. I do not attempt to articulate what a user should do in this instance. As discussed in the introductory chapter, I have set the issue of track record aside for the course of this dissertation. There are instances in which track record is unavailable or not easily interpreted, and as such, it is worthwhile to appraise the justificatory value of explanations on their own. With that said, back to the topic at hand.

Feature salience explanations, also called salience or sensitivity mapping, is another common AI explainability strategy in which the factors that most heavily influenced an AI system's output are highlighted for the explainee (Dabkowski & Gal, 2017; Simonyan et al., 2013; Zeiler & Fergus, 2014; Baehrens et al., 2010). Saliency mapping is often used to explain image classifications by identifying the pixels in an image that were most important in deriving the system's image classification. For example, an image classification system that identifies skin blemishes as cancerous or non-cancerous highlights the portions of the image that most strongly influenced the diagnosis (Esteva et al., 2017). The AI system has no concept of skin blemish or cancer. It sees only pixel color and proximity, but by highlighting areas of an image, the AI system's image classification process is framed in a way that is understandable to humans. A user might not be able to understand the algorithmic process by which an AI system classifies an image, but by the AI system pointing out parts of the picture the system finds most important, the user can compare the system's reasoning to her own background understanding of how the image classification should work. For example, Eric Topol (2019) describes a case in which an AI diagnostic system learned to identify pictures of cancerous skin blemishes based on the presence of a ruler or caliper in the image. In the data on which the system was trained, blemishes that dermatologists suspected to be cancerous were more often photographed with a measuring instrument than those that were not. A lay human user might not know how to tell the difference between a cancerous and non-cancerous skin blemish, but assuming she does know the diagnosis should have something to do with the blemish in question, the user should be more confident in the outputs of a classification system that highlights pixels corresponding to the blemish and not to the ruler also present in the picture.

Finally, *counterfactual explanation* techniques tell users what minimum change in input conditions would have resulted in an AI system outputting a different decision or recommendation (Watcher et al., 2018; Fernandez et. al., 2020; Halpern & Pearl, 2005). Like saliency mapping, counterfactuals provide insight into which factors most heavily influence a decision, though counterfactual explanations go a step further than salience maps in helping explanation recipients see how those factors influence decisions. For example, a medical treatment planning system might indicate that a specific cancer treatment recommendation would have been different had a patient's white blood cell count been a little lower or had the patient responded well to other treatments in the past. In some instances a user might also manipulate relevant variables to see how the system

responds, which allows the user to build a causal mental model of how the system derives its outputs – e.g. changes to input X cause changes to the outputs in such and such a way. In general, if that mental model coheres well with a person's mental model of how these factors influence each other in the real world, then she may be more justified in believing the AI system's outputs.

Note that AI systems do not actually reason causally; to build on Alex London's (2019) example, an AI system does not, for instance, make predictions about a patient's risk of death from pneumonia by querying its understanding of the biomechanics of respiration and the possible complications thereto posed by infection and any preexisting respiratory conditions. For instance, asthma is a respiratory condition characterized by constricted bronchioles. The constriction makes it difficult for sufferers to expel infected phlegm by coughing. Consequently, pneumonia infections developed by asthma sufferers are more likely to become severe thus posing a higher risk of death from pneumonia to those patients. If an AI system offers a counterfactual explanation for its classification of an asthmatic being at high risk of death – "had the patient not suffered from asthma the risk score would have been lower" – the explanation may seem to imply that the AI system employed some causal reasoning about the impact of asthma on pneumonia recovery to make its prediction. But this is not the case. Any such causal reasoning a user perceives AI system's counterfactual explanation to demonstrate will have emerged from the data sets on which the AI system was trained; the risk score will be lower simply because the masses of data on which the AI system was trained shows that non-asthmatics do recover more often than asthmatics. That said, by framing AI reasoning processes in causal terms, counterfactual explanations still make AI reasoning processes more easily compared to the background knowledge a human user might hold about relevant causal relationships.

In sum, AI explainability techniques deviate from the internal processes by which AI systems derive conclusions so as to frame those processes in more familiar terms. AI systems do not themselves reason like humans naturally tend, but explanations like prototypes, saliency maps, and counterfactuals can be more easily appraised by humans for coherence with their own beliefs and background knowledge. In this way, AI users might be able to use AI explanations to gain some internal epistemic justification for their belief in AI outputs even in the absence of system track record.

There is, however, the glaring complaint against coherentism that I set aside earlier. If one's preexisting beliefs are false, then coherence with those beliefs is not an epistemic virtue; coherence with false beliefs is not an indication that a new belief is epistemically well-founded or likely to be true. This complaint has a seemingly troublesome implication for the justificatory value of explanations. To illustrate, imagine an astrologer diagnoses Jack with cancer and offers a treatment recommendation based on her appraisal of star charts and astrological events. The astrologer explains her reasoning to Jack who is an amateur astrologer himself. Jack believes that medical diagnoses are most likely to be true, and that treatment plans are most likely to be effective, when reasoned to via

astrological investigative methodologies. Jack is also familiar with the astrological theories and methodologies the astrologer described in her explanation. Does the explanation that the astrologer provided hold any justificatory value for Jack? According to the definition I have provided, it does. The astrologer's explanation is comprehendible to Jack, and it provides Jack with information which, via strong coherence with his existing background beliefs, gives him internal epistemic justification for his belief in the astrologer's diagnosis and recommendation. But surely it seems wrong that an explanation for a cancer diagnosis based on star charts and planet positions would be justificatorally valuable. Afterall, the point of justificatory value is that a justificatorally valuable explanation provides the recipient epistemic justification for her beliefs, and presumably stronger epistemic justification should be held in those beliefs which are indeed epistemically well-founded and likely to be true. (Let us assume that medical diagnoses based on astrological phenomena are not).

Perhaps my definition of justificatory value needs to be updated to include some requirement for external epistemic justification in addition to internal justification. For instance, I might say that a justificatorally valuable explanation is one that (a) provides internal justification via coherence with background knowledge and beliefs and (b) points the recipient in the right direction – that is, toward believing claims that are likely to be true and away from those which are not. Another option is to say that a justificatorally valuable explanation is one that (a) provides internal justification via coherence with background knowledge and beliefs and (b) is a post hoc rearticulation of an attempt to employ "epistemically legitimate" or epistemically sound methods of reasoning. For instance, a justificatorally valuable explanation for a diagnosis would not be able to reference astrological methodologies (we have assumed these are not epistemically sound methodologies), but it could be based in biomedical science. However, we now face a new problem. We must now define what epistemically "legitimate" or sound methods are. For instance, what is it, exactly, that differentiates astrology from science? A philosophical skeptic would argue that there is no way to tell; we might all (scientific experts and laymen alike) be like astrologers in the sense that we are deeply embroiled in wildly false yet elaborate and internally consistent webs of belief. And, if such is the case, we have no way of knowing![3] I am not going to solve this problem here, let alone provide anything resembling a satisfactory overview of those who have tried. For the time being I am happy to acknowledge that my original definition of justificatory value has a hole. That said, there are a couple points worth making which may ameliorate the situation.

First, note that if we take the skeptic's viewpoint seriously, we have a much bigger, more fundamental problem on our hands than defining justificatory value. It is impossible to tell whether anything we believe is true, and no amount of explaining will ever fix this.[4] But for the sake of philosophical progress within the experienced reality that we do all share, let us assume that it is

[3] Gilbert Harman's (1973: 5) brain-in-a-vat thought experiment – articulated more famously by Hilary Putnam (1981: 1-21) – comes to mind.
[4] Comesaña & Klein (2019) provide a thorough overview of such arguments for philosophical skepticism.

possible to differentiate with some degree of success between areas of expertise or ways of knowing that are more epistemically sound and those which are less so. After all, it is generally acknowledged that the sciences have something epistemically desirable that astrology lacks. At this time, however, I will not attempt to specify exactly what that something is.

This leads to my second point. I posit that instances in which an explanation recipient is embroiled in an elaborate and internally coherent, though radically false, set of beliefs – and where those beliefs are relevant to the appraisal of an explanation – are not the norm. For the most part, when discussing the justificatory value of explanations, we are not, for instance, interested in astrologer explanation recipients who think they know something about medical diagnostics. (Again, if such cases are the norm, we have a much bigger problem on our hands than the definition of justificatory value; why are so many people committed to epistemically dubious ways of knowing?). Rather, we are concerned with instances in which the explanation recipients have varying degrees of background knowledge in the "legitimate" domain of expertise to which the explanation pertains, such as the extent of one's medical knowledge when appraising an explanation for a medical diagnosis or one's proficiency in economic theory when seeking investment advice. So going forward, I refer only to these cases. More so, I am specifically interested in epistemically imbalanced relationships – situations in which the explanation recipient lacks background knowledge in the relevant domain of expertise relative to the explainer. In epistemically imbalanced relationships such as human expert-novice relationships and many user-AI interactions, the explainer utilizes greater knowledge and superior cognitive capacities to inform her/its reasoning than the explanation recipient has at her disposal. Consequently, given the relatively limited epistemic resources available to novices/users, there will likely be gaps in novice/user background knowledge that correspond to key points in an expert's reasoning process which will either be ignored or mislead the novice's evaluation of the expert's claim. Accordingly, the maximum potential justificatory value of an explanation – that is, the explanation's ability to help the recipient make an epistemically well-informed decision about what to believe – is capped by the background knowledge and cognitive capacities of the explanation recipient. The next section illustrates this phenomenon and its implications for the justificatory value of expert-to-novice and AI-to-user explanations.

2. CONTINUUMS OF JUSTIFICATORY VALUE

Irrespective of how detailed or reflective an explanation is, an explanation cannot be justificatorally valuable to an explainee if the explainee is not endowed with relevant background knowledge or epistemically well-grounded beliefs against which to assess its content. Inversely, the more well-versed a person is in a domain of knowledge, the greater the potential justificatory value an explanation pertaining to that domain can hold for them. The result is a *continuum of potential justificatory value* ranging from explanations of higher maximum potential justificatory value – those

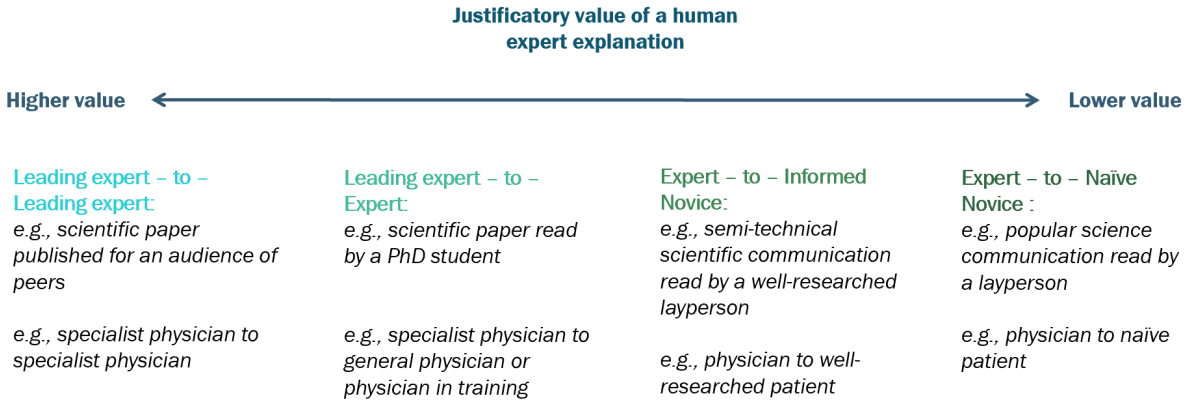
received by people with extensive knowledge and well-honed reasoning skills in a domain (experts) – to explanations with lower maximum potential justificatory value – those received by people with little background knowledge in the relevant domain (novices or laypeople).

I speak in terms of maximum potential justificatory value instead of realized justificatory value because it is possible for an explanation to be offered that contains less information than the recipient would be able to understand, and which would therefore underutilize the recipient's background knowledge in the domain. Such would be the case, for instance, if a physician excessively "dumbs down" the explanation of her reasoning for a very well-informed patient. Such an explanation does not achieve the maximum justificatory value it could, given the cognitive capacities and background knowledge held by the recipient.

Accordingly, leading-expert to leading-expert explanations are the best-case scenario for the justificatory value of an explanation according to coherence theories of justification (See [figure 2.1]). When an expert explains her reasoning to another expert in her field or constructs a post hoc and partial account of her reasoning in the form of a peer reviewed scientific paper, a high degree of technical knowledge on the part of the recipient can be taken for granted. Accounts exchanged between experts can be complex and can reference domain specific knowledge while remaining cognitively accessible to the expert recipient. The greater complexity and detail of an explanation affords more opportunity for the recipient to compare the contents of the explanation with her own knowledge and background beliefs about the topic, and the expert's high proportion of true and well-founded beliefs means that coherence with those beliefs buttresses one's justification for believing what the explainer has to say. For example, imagine that a physician offers an explanation for a diagnosis and treatment plan to a patient named Jan. Jan is a final year medical student. Knowing this about Jan, the physician offers Jan a detailed explanation for her treatment recommendation. The physician discusses Jan's diagnostic tests in detail, and outlines how she used those tests to reason to the given diagnosis while discounting other diagnoses (ones Jan learned about in training) that can present with similar symptoms. More so, the physician accounts for the expected efficacy of the treatment plan and possible side effects in terms of both clinical trial data and the molecular mechanisms by which various prescribed medications will take effect. Because of her medical training and her well-developed background knowledge in biomedical science, Jan is able to appraise the explanation for possible errors in reasoning and false premises. Note, however, that Jan's appraisal would be even more well-informed, and therefore the explanation could be of even greater justificatory value to Jan, if she were further along in her medical training; the explanation could be more detailed and Jan would be even more likely to identify false premises or errors in reasoning.

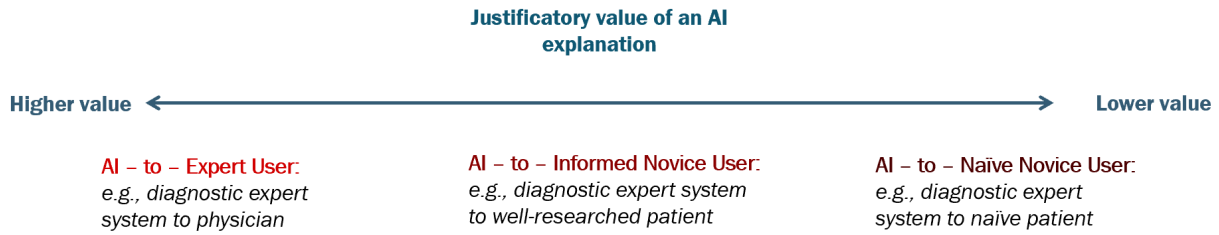
On the other hand, when an expert explains her reasoning to a novice (as when a scientist communicates findings to the public, or a physician to a lay patient), the novice explainee has less relevant knowledge about the field with which to compare the expert's reconstructed reasoning than a leading expert or expert-in-training. For example, imagine that a physician offers an explanation for a

diagnosis and treatment plan to a patient, Anne. Anne is a tax accountant whose familiarity with medical sciences ends with her high school biology class. In comparison to the explanations offered to Jan, any biomedical explanations that Anne receives from our physician must be simplified to remain cognitively accessible to Anne. Not only does the simplification of the explanation reduce the information present for Anne to appraise, thus limiting the insight Anne can draw from the explanation, but given Anne's limited background knowledge in the biomedical sciences, she is less likely than Jan to identify errors in reasoning or false premises in the information the explanation does contain. The more novice the explanation recipient is, the more the recipient must take the expert's word not only about the end finding, but about the truth or well-foundedness of the premises used to account for that finding as well. In this way, the maximum potential justificatory value of an explanation is limited by the explanation recipient's background beliefs and her cognitive capacities.



[Figure 2.1] The maximum potential justificatory value of an explanation is capped by the cognitive capacities and relevant background knowledge held by the explanation recipient.

The justificatory value of a given explanation will vary depending upon who is receiving the explanation, and this will be the case whether the explanation source is human or AI. As in human expert-novice relationships, the limit is not on the ability of an AI system to produce highly detailed and reflective explanations – an AI system could output a highly accurate account of its computational processes – but on the explanation recipient's capacity to comprehend the explanation and on the extent of the recipient's relevant background knowledge. Accordingly, the same continuum of potential justificatory value applies to AI explanations and to explanations offered by humans [figure 2.2]. AI-to-Expert explanations, such as explanations offered by AI medical diagnostic systems to physicians or AI financial planning assistants to human investment bankers, can hold a higher maximum potential justificatory value than those offered to lay users such as explanations offered by medical diagnostic systems to lay patients or AI financial planning assistants to general users.



[Figure 2.2] The maximum potential justificatory value of explanations offered by AI systems fall along the same continuum as analogous explanations offered by humans.

The continuum of justificatory value becomes most interesting when we focus on the right side of the spectrum where the maximum potential justificatory value of an explanation approaches zero. I posit the perhaps uncomfortable conclusion that a significant proportion of explanations offered by experts to novices and by AI systems to lay users fall in this space, where the potential justificatory value of an explanation is capped at little to none. In other words, many expert-to-novice and AI-to-lay user explanations do not contain information that the explainee can use to help epistemically justify her belief in or acceptance of claims/outputs. This is not due to the explainer being unable to produce an explanation that accounts for her claims, but to the explainee not being able to comprehend a more valuable explanation and/or not possessing the relevant background beliefs for coherence with that knowledge to provide a degree of internal justification for an epistemically well-founded belief.

Consider, for example, an AI system that explains its diagnosis of malignant tumors in breast scans by highlighting the pixels most relevant to its decisions (a saliency mapping explanation technique). Based on such an explanation a naive lay patient might determine that the system's reasoning processes responds to something observed in the image, but the patient does not have the relevant medical knowledge to appraise whether the system is responding in the "right" way – in a way such that its identification of a malignant tumor is an accurate evaluation of the patient's health. The same would be true, however, if a human physician were to show a patient the areas on her scan that indicate the presence of malignant tumors. In both the human expert case and the AI case, the explanation might suggest to the lay recipient that the expert/AI could provide a more justificatorally valuable explanation to a more epistemically well-equipped individual if required, but the ability of the explanation to help the novice independently acquire well-placed internal epistemic justification for her belief in the claim is very low. On the other hand, if either the AI system or the human physician were to show a second physician the same scan, the second physician would be able to see that the AI system/first physician is responding to something it/she observed in the scan, and by referencing her own medical experience and background knowledge, the second physician would be able to appraise whether the features highlighted in the image are relevant to and support a cancer

diagnosis. For the physician explanation recipient, the justificatory value of the AI system's explanation lies further to the left on the spectrum that for the lay patient [figure 2.2].

The key takeaway here is that if the limiting factor on the justificatory value of an explanation is that the extant beliefs against which the contents of an explanation are compared are often flawed or lacking, then the source of the explanation, whether it be human or machine, is not our greatest concern. Rather, it is that the maximum potential justificatory value of an explanation is limited primarily by the cognitive capacities and background knowledge of the explanation recipient, and as such, when an explainee is a lay user or naïve novice, that maximum potential justificatory value will necessarily be quite low. This will be the case whether the explanation source is a human expert or AI, and it seems the only way to raise the cap on the potential justificatory value of an explanation in either case would be for the explainee to become an expert in her own right. But then what would have been the point of seeking expert/AI advice in the first place? We are right back where we started in Chapter 1 grappling with Meno's paradox; it seems one must be an expert to make an epistemically informative evaluation of expert or AI explanations.

More so, we have gained some insight into the debate around the doing-saying mismatch in scientific papers; even if we decide contra Medawar that it is just fine for the published justifications for scientific findings to stray from the original discovery process, the fact remains that scientific papers remain impenetrable to most people and that lay accessible reports of scientific discoveries contain little information that readers can use to help epistemically justify their belief in scientific claims. Therefore, for all but the most expert readers in the relevant domain, the justificatory value of scientific papers is substantially limited. So, like Medawar, I am skeptical about the justificatory value of scientific papers, but for a very different reason. While Medawar's skepticism is a response to how scientific papers abstract from original scientific discovery processes, my skepticism is a response to the cognitive limitations of those who might read the papers, not the papers' content.

3. THE JUSTIFICATORY VALUE OF COMPLETE FABRICATIONS

Before I conclude this chapter, I would like to comment briefly on the justificatory value of lies – that is, of fully non-V2-veridical explanations, cases in which an explainer completely fabricates an explanation with the goal of directing the recipient to a desired belief. There certainly is some reason to prefer explanations that bear some connection to the reasoning process which is supposedly being recounted; as I discussed in Chapter 1,reflectivity can contribute to an explanation's justificatory value by providing the recipients with indications of the safety and sensitivity of the explainer's reasoning process. That said, I also noted that there are cases in which dramatic abstraction in an explanation's reflectivity can boost justificatory value. Such is the case when an expert arrives at a conclusion by intuition but offers an explanation other than "it felt right", or where AI system explanations need to be heavily modified to cater to human understanding. So, while

reflectivity can help improve an explanation's justificatory value, it is not the case that a more V2-veridical explanation should necessarily be preferred. I posit that if there is a universal problem with explanations that are completely fabricated accounts of reasoning, it is not that these explanations inherently lack justificatory value. If anything, the common problem fabricated explanations present is a moral one. For instance, we might take Immanuel Kant's (1993, 1996, 1997) view that lying somehow robs people of human dignity by infringing on their rational autonomy. So, according to Kant, to tell a lie is never morally permissible regardless of its epistemic outcome. But whether one ought to agree fully with Kant's absolute moral prohibition against lying is another discussion entirely.[5] Here I am interested in the justificatory value of fabricated explanations – that is, the explanation's ability to provide recipients with information they can use to help epistemically justify their belief in claims, outputs, or recommendation.

Imagine a case in which an astrologer references star charts and astrological phenomena to diagnose Jan's (the final year medical student) and Anne's (the tax accountant) symptoms and to derive a treatment plan for each patient. The astrologer knows Jan and Anne both believe treatment plans are most likely to be effective when reasoned to in accordance with the knowledge and methodologies of biomedical science. As such, the astrologer fabricates explanations for Jan and Anne that reference biomedical terminology and mechanisms of disease. These explanations are non-V2-veridical in the fullest sense; they in no way reflect the reasoning processes the astrologer employed to derive her diagnoses.

Is the astrologer's explanation of any justificatory value to either Jan or Anne? I argue that it is for Jan, and might be for Anne. If the astrologer's explanation for her treatment plan is good (in the sense that it is V1-verdical and, as such, supports her claim well) then it is a good explanation irrespective of how it relates to the astrologer's original reasoning processes; in which case, the greater the recipient's biomedical expertise (Jan's more so than Anne's) the better able the recipient will be to see that the explanation does indeed support the astrologer's claims well. It may happen, of course, that in attempting to fabricate a "science-sounding" explanation the astrologer makes mistakes. But again, the greater the recipient's expertise in the biomedical sciences, the more likely she will be to note the explanation's flaws which would in turn lead her to doubt the astrologer's recommendations. Final year medical student, Jan, would likely be quick to note that something is off. Tax accountant, Anne, on the other hand, could be more readily led on. As such, the explanation is more justificatorally valuable to Jan than to Anne. Indeed, if Anne's biomedical background knowledge is sufficiently lacking, the explanation may be of no justificatory value to her at all.

So, a completely fabricated explanation can be justificatorally valuable, and, in accordance with the continuums of justificatory value I described in Section 2, just how justificatorally valuable the fabrication will be is primarily constrained by the recipient's cognitive resources. More

[5] And for those interested in that discussion, *inter alia* see Carson (2010), Mahon (2006, 2009), and MacIntyre (1995).

importantly, this will be the case regardless of the explanation source. If the explanation were instead offered by an AI-system, it would still be unreflective of the complex statistical processes by which the system derived its output. The astrologer is interchangeable with an AI-diagnostic system in the case described above. Similarly, if the astrologer's explanation were instead offered by a physician (from whom the explanation might be more V2-veridical), Anne/Jan would still be just as (un)able to use the content contained in the information to appraise the veracity or epistemic well-foundedness of the diagnosis via coherence with her own background knowledge and beliefs. So, the real concern is not about the explanation source (human or artifact, legitimate scientist or charlatan) or even about whether the explanation is reflective or fabricated. The real concern is that an explanation will inherently be of lower justificatory value the less relevant background knowledge a recipient has in the domain of expertise to which the explanation pertains.

4. CONCLUSION: DOWN WITH EXPLANATION

This chapter has proceeded in two main steps. In Section 1, I first described how a person may acquire a degree of internal epistemic justification for her belief in a claim or output by evaluating how well the accompanying post hoc and partial account of reasoning coheres with her existing background knowledge. In Sections 2 and 3, I then observed that in epistemically imbalanced relationships like those characteristic of novice-expert and AI user relationships, the justificatory value is limited by capacities of the recipient and noted that this is true whether the explanation source is human or AI. This observation led to the overarching conclusion that in heavily epistemically imbalanced relationships in which the novice/user has little to no relevant background knowledge, explanations will often be of minimal justificatory value; they will not contain much information that can help a recipient make an epistemically well-informed decision to (dis)believe expert claims and AI outputs.

Explanations do, of course, have other functions beyond providing epistemic justification for claims. For instance, in patient-physician relationships explanations provide opportunity for patients to ask questions, help patients to psychologically and emotionally prepare for the risks and consequences of treatment, and provide indications as to whether the values and priorities underpinning physician decisions and recommendations align with the patient's own (Manson, 2010). These other functions should not be overlooked, but the epistemic concern still remains; for lay users and novices, explanations provide little epistemic basis for believing the outputs and claims of experts and AI systems alike. As such, we must not be very hopeful that AI or expert explanations can improve the epistemic security of user-AI or novice-expert communications.

I understand that this is an overwhelmingly negative conclusion, but I will end on a more positive note. The post hoc and partial explanations offered by human experts to novices are subject to the same restrictions in justificatory value as the post hoc and partial explanations AI systems offer to

users. Both are constrained by the limited background knowledge and cognitive capacities of the epistemically disadvantaged recipient. Yet, laypeople daily defer to human experts – to physicians, electricians, accountants, plumbers, mechanics, financial advisors, and lawyers – to help solve their problems while acquiring only a minimal grasp of the experts' methods. If we are so concerned with accepting AI outputs without confirmatory explanation, then we ought to be more concerned about relying so heavily on human experts as well. On the other hand, if we assume that novice reliance on expert advice is not totally unwarranted, then there must be some basis for deciding when to take an expert at her word and the same could be true for AI as well. In these first two chapters I have demonstrated, however, that looking to AI explainability strategies to help underpin epistemically justified beliefs is barking up the wrong tree.

Where explanations fall short as a source of epistemic justification, I propose that we instead turn the discussion to trust. For the time being, trust can generally be understood as the willingness to rely on or believe an external entity without guarantee of how that trustee will perform or without being able to verify the trustee's claims for oneself.[6] For example, as lay patients, when we visit a physician, we do not believe what the physician has to say because we were able to verify the content of the physician's claims using whatever explanations were provided. Rather, we believe that, in general, we would do well to heed the physician's advice, perhaps because we understand that the physician has completed required medical training, or because physicians adhere to general norms of communication or conduct which provide us with some default basis for believing what others tell us.[7] Overall, we have reason to trust the physician as a source of epistemically well-founded medical advice, and it is this trust that gives us some degree of epistemic justification for believing the physician which, in turn, provides a degree of epistemic justification for believing the physician's claims.

Interestingly, even where explanations cannot help novices and users to evaluate AI/expert outputs directly, explanations might still indirectly contribute to justified beliefs by substantiating a degree of epistemic trust in experts or AI systems. For example, Alvin Goldman (2001) describes how the dialectical ease with which an expert presents her explanations to novices provides an indication of the expert's comfort and familiarity with the topic she is discussing. This, in turn, can be taken as an indication of speaker competence (and a competent speaker is more likely to present true or epistemically well-founded claims than an incompetent one). That being said, the same relationship is unlikely to hold for AI systems. For humans, cognitive limitations to our creative capacities make it much easier to recount information and reasoning processes with which one is familiar than to fabricate convincingly detailed accounts on the spot. However, AI systems are not so cognitively

[6] Karen Jones (1998) provides a concise overview of the trust literature in which she identifies acceptance of risk in relying on others (or in other words, a lack of guarantee of trustee performance) as one key and widely accepted feature of trust.

[7] See Philip Nickel (2013) on norm-based entitlement for justified belief in testimony. I attend further to common norms of conduct and communication as a basis for trust in AI and in Experts in Chapters 3 and 5.

limited and may find fabricating explanations equally (or more) efficient for attracting user trust as recounting truth (Irving & Askell, 2019; Irving et al., 2018).

So overall, explanations are limited both in their ability to directly provide a degree of epistemic justification in expert claims and in their ability to indirectly underpin a degree of epistemic trust in the explanation provider (especially when the provider is an AI system). As such, the foundation of epistemic trust in experts and AI systems warrants further investigation. In the following chapters I address questions such as "What makes a trustee epistemically trustworthy?", "How is epistemic trustworthiness indicated to trustors?", "How are foundations of trust built?", and "Can we build similar foundations for trust in AI?"

PART II

TRUST

3

Epistemic Trust in Experts and AI

In previous chapters I have investigated how a person might acquire a degree of epistemic justification for believing the outputs of experts and AI systems. I have argued, however, that in epistemically imbalanced relationships, like novice-expert and user-AI relationships, the epistemically disadvantaged party is often ill-equipped to justify her belief in expert/AI outputs based on the content of the output or an explanation of the process by which it was derived. But this does not mean that we must be resigned to novice beliefs in expert/AI claims always having little epistemic basis. Where the epistemically imbalanced nature of the relationship precludes direct epistemic appraisals of expert claims, another option is to appraise the epistemic trustworthiness of the claim's source as opposed to the claim itself. Within the context of epistemology of testimony, Richard Moran (2006) describes this as a switch in focus from pursuing justified belief in testimony itself to pursuing justified belief in the testifier. The idea is that if a person has reason to believe that a speaker is an epistemically trustworthy source of information regarding topic X, then she holds a degree of epistemic justification for believing the speaker regarding topic X, and, in turn, holds a degree of epistemic justification for believing the speaker's claims regarding topic X. The next three chapters investigate the foundations of epistemic trust in human experts and AI systems.

First, I would like to acknowledge that some will balk at my use of trust terminology with reference to AI systems. Trust, it is commonly argued, is a term reserved for a specific class of actors that are capable of sincerity and deception, holding goodwill towards others, or taking responsibility for their actions or claims (*inter alia* see Baier, 1994; Holton, 1994; Jones, 1996; O'Neill, 2002). Computational machines like AI systems (or any computational or mechanical instrument for that

matter) do not qualify as proper objects of trust and therefore instead ought to be spoken of in terms of reliance. Chapter 5 is dedicated to addressing these concerns. For the time being, trust can generally be understood as the willingness to rely on or believe an external entity without guarantee of how that trustee will perform or without being able to verify the trustee's claims for oneself.

At this time, I am most interested in a particular species of trust called epistemic trust. This dissertation is concerned with how people can acquire some epistemic justification for their belief in expert claims and AI outputs, and epistemic trust specifically describes one's willingness to believe a trustee's claims based on one's perception of the trustee's epistemic trustworthiness. Epistemic trustworthiness refers to the likelihood of the trustee producing true or epistemically well-founded outputs. Epistemic trust can be contrasted with practical trust which is trust in an external entity to do certain things like trusting a neighbor to collect your mail without violating your privacy. Practical trust can also overlap with epistemic trust. For example, I may trust my physician to keep up with the latest medical research (practical trust) which, in turn, informs my belief that she will likely provide me epistemically well-founded medical advice (epistemic trust). Unless specified otherwise, all references to trust going forward should be assumed to refer to epistemic trust.

The next three chapters on trust in experts and AI are divided as follows. First, the remainder of this chapter investigates the nature of epistemic trust in human experts and in AI by looking into what factors can be expected to, and should most strongly, influence appraisals of the epistemic trustworthiness of experts and AI. Most philosophical discussions of epistemic trust in experts pertain to what I will call, *idiosyncratic trust* – trust based on things a novice knows about an individual expert such as her track record, training, and motivations. However, I caution that discussions that frame challenges to novice/user trust in experts/AI primarily as issues of idiosyncratic trust can misdirect practical conversations about evaluating trustworthiness and attracting user/novice trust. Novices very often cannot mount meaningful appraisals of the trustworthiness of individual experts based on their idiosyncratic features. Rather, on a daily basis, novices quite successfully defer to the recommendations and services of experts while knowing very little about the individual experts on whom they rely; patients are generally well served by relying upon physicians to address their medical needs, car owners upon mechanics, homeowners upon plumbers, citizens upon tax accountants, and so forth. Of course, there are bad apples, quacks, and posers, and even the most well-intentioned and competent experts can make mistakes, but for the most part, novice-expert alliances prove fruitful with the basis for the novice's trust in the expert not extending far beyond her belief that the expert is one of a type, a person who supposedly has high levels of specialized knowledge and skill in a particular domain.

Accordingly, in this chapter I describe the general success of novice trust in experts as the result of *systemic trust* as opposed to idiosyncratic trust, where systemic trust is trust held in an expert grounded in a novice's belief that an expert is a member and/or product of an expert community and larger social-epistemic environment that influences the expert's behavior and performance. In Section

1, I provide a more detailed description of systemic trust in relation to idiosyncratic trust. In Section 2, I illustrate that systemic trust best accounts for novice willingness to defer to experts, and I argue that discussions about grounding novice trust in experts should be framed as an issue of systemic trust. In Section 3, I argue that systemic trust pertains to user trust in AI in much the same way as it pertains to novice trust in human experts. Section 4 concludes this chapter. Chapter 4 then briefly interjects to address concerns regarding the use of trust terminology to describe attitudes held toward non-human entities like AI technologies and social-epistemic system, and Chapter 5 discusses how foundations for systemic trust in both human experts and in AI-enabled expert systems can be built and demonstrated to trustees. Overall, my point is that for the purpose of grounding epistemic trust in AI, systemic factors like policy and community norms play a far larger role than idiosyncratic factors like the explainability of individual AI systems.

1. WHAT IS SYSTEMIC TRUST?

When you go to see a physician, you might tell a friend, "I'm going to see a doctor." You are not likely to say, "I'm going to see a *trustworthy* doctor". The specification of the physician's trustworthiness feels redundant for a couple of reasons. First, it would be very odd to knowingly choose to consult an untrustworthy physician, so unless you have a history of doing so (an odd habit), you would not specify otherwise. Second, a physician's trustworthiness is largely assumed. Something about the physician being a physician already implies that, in general, one would do well to hear her medical recommendations. Among other things, physicians undergo standard training, are held to ethical standards, are subject to performance and ethical reviews, and can be disbarred if they perform poorly or breach professional standards. Such factors contribute to what I call a patient's *systemic trust* in a physician. While *idiosyncratic trust* is trust based on things a novice believes about an individual expert such as the expert's character, past performance, intentions, credentials, or personal values, systemic trust is the kind of trust a novice would hold in an expert because the expert is embedded in a network of institutions, practices, policies, and norms of conduct and communication – hereafter called *systemic factors* – that influence the trustee's behavior.

One might point out, of course, that how an expert is embedded in a larger social-epistemic system is also an idiosyncratic feature. For example, that a practicing physician is a member of the medical profession or certified by a specialist medical board helps to describe the individual physician. However, such idiosyncratic features as community membership, educational background and credentials are only meaningful in virtue of the capacity of the larger social-epistemic systems in which experts are embedded to influence individual expert conduct and behavior. I detail this point further in section 2.

I am not the first to note that idiosyncratic features often do not explain (or do not fully explain) novice trust in expert testimony. Philip Nickel (2013) makes a similar observation in his

paper differentiating between assurance-based and norm-based entitlement for justified belief in testimony. In cases of assurance-based entitlement, it is argued that one's justification for belief in an instance of testimony lies in the fact that in making a claim, the speaker accepts some responsibility for the claim thereby providing assurance of its content (Moran, 2006; Hinchman, 2005). However, Nickel points out that especially in cases of expert or specialist testimony a hearer likely knows very little about the speaker – that is, too little for the speaker's own assurance of the quality of her testimony to provide the hearer with good reason to believe it. For example, Nickel observes that "when considering whether to accept the testimony of a scientist I do not know at all, and about whom I have only limited contextual information, it does not really help me to know that this person offers me a personal guarantee that her statements are true" (216). Rather, Nickel continues, "in this context norm-based entitlements are in a better position to respond to critical doubts than epistemic entitlements based on speaker assurances" (216). Norm-based entitlement for belief exists "where general epistemic norms of behavior and communication that publicly obtain," and, Nickel points out, "testimony in science is based on norms that are explicitly discussed, propounded and enforced by the community" (215).

In a line of reasoning similar to Nickel's, I also argue that idiosyncratic factors (those things one might know about an individual expert) do not provide a great foundation for novice trust in expert testimony more generally. Novices simply do not know much about the individual experts they consult. I defend this point in greater detail in Section 2. My notion of systemic trust, however, is broader than Nickel's idea of norm-based entitlement. In addition to norms of communication and conduct, I also consider how explicit policies, rules, regulations, and institutional structures underpin systemic trust in experts as well.

In this respect, my notion of systemic trust bears some resemblance to Frederic Bouchard's (2016) account of "institutional trust". With regard to lay trust in scientific expertise, Bouchard writes, "It's not that scientific experts are necessarily right or trustworthy. It's that scientific experts' actions and claims are checked by institutions that have useful critical epistemic values and ideals. Scientific experts are not trustworthy because they are objective and right about the truth of the universe, but because they've been given credentials by a system that is skeptical and critical about expertise" (599). Bouchard is arguing contra Goldman (2001) that novices will often have little reason to trust individual experts because, as they are not experts themselves, they do not have the knowledge or skill necessary to evaluate the putative expert's expertise. More so, Bouchard continues, it will sometimes be the case that novices have historically rooted reasons to actively distrust individual experts. For example, following the Tuskegee syphilis trials, African Americans have historical experience as a reason to distrust medical researchers claiming to work for African American benefit. Medical research policies and ethical research requirements have since changed with the aim of preventing such medical misconduct, though, Bouchard argues, people are not so quick to change; individual researchers today, if not so constrained, would repeat past atrocities. Therefore, Bouchard argues,

rational novice trust in science is only via trust in the institutions that check expert behavior – which Bouchard calls "institutional trust" – as opposed to trust in individual experts.

My own notion of "systemic trust" is along similar lines to Bouchard's notion of "institutional trust" in that systemic trust is not based in an individual expert's idiosyncratic characteristics but on the institutions and larger systemic infrastructures in which the expert is embedded. However, I modify Bouchard's account in a few key ways. First, I do not limit my discussion to lay trust in science. My discussion also includes other applications of expertise such as in medicine, law, engineering, or architecture that are influenced by the institutions of science to varying degrees. Second, I see systemic trust as a kind of trust that can be held in individuals, as opposed to trust held in systems or institutions while individuals are actively distrusted. If a novice bases her trust in an expert on assumptions about the systemic backing that regulates or influences the expert's behavior, then to trust an expert is largely about trusting the system in which the expert is embedded. Similarly, to trust the system is to have reason to trust the expert. So, to clarify, idiosyncratic trust is the kind of trust a novice holds in an individual expert based on an expert's *idiosyncratic factors* such as an expert's unique track record, characteristics, motivations and so forth. Systemic trust is also trust held in an individual expert but is based on the networks of *systemic factors* in which the expert is embedded as opposed to the expert's individual characteristics. The distinction between idiosyncratic trust and systemic trust in individuals is useful for structuring discussion, but it is important to note that the delineation is not perfect. Some idiosyncratic factors only provide evidence of individual expert trustworthiness if systemic trust is presupposed. Such is the case, for instance, when reviewing an expert's credentials, certifications, educational background, or expert community membership. Each of these factors only serves as an indication of expert trustworthiness if it is assumed also that the educational and certifying institutions do indeed hold experts to high standards, that is to say that the "system" is working as it ought to be. In Chapter 5, I discuss such systems as *epistemically well-functioning systems*.

Third – and this is what makes systemic trust 'systemic' as opposed to 'institutional' – recall that in addition to institutional structures, rules, and policies, I include factors like cultural values and Nickel's norms of conduct and communication among the systemic factors that influence individual expert behavior. The systemic factors that underpin systemic trust in an expert roughly fall into two categories: extrinsic factors and intrinsic factors (Baum, 2017; Le Grand, 2003: 53). Extrinsic factors are measures externally imposed on experts and expert communities to induce experts to behave in a certain way and maintain certain standards regardless of their desire to do so. Extrinsic factors include policy, certification requirements, monetary incentives, education standards, and audits. On the other hand, intrinsic factors such as cultural values and norms of conduct and communication, emerge within communities of experts and cultivate within experts a desire to behave a certain way and to maintain certain standards.

The development of norms is a complicated process which I discuss in more depth in Chapter 5. The important point here is that the evolution of social norms and values is, in part, facilitated by the implementation of extrinsic measures like education requirements or rules of conduct that set a precedent for and catalyze new normal behavior (Coleman, 1994; Seger, 2022). In this way, the internal motivations of experts, which are idiosyncratic features of those experts, are often linked to the systemic factors. Furthermore, one's reason for believing that an expert is self-motivated in some way is often based on one's knowledge of the system in which she is embedded. For example, I assume my physician keeps up with modern medical research not because I know she is a particularly studious individual, but because I know she is a licensed practicing physician who went through a rigorous training program that would have required her to develop studious habits. In this way, even reasoning about idiosyncratic factors often requires background reasoning about systemic factors.

A key feature of systemic trust is that individual experts at the same node in a network of systemic factors are interchangeable. For example, if a patient's regular general practitioner (GP) goes on holiday and the medical practice schedules a second physician to cover the regular GP's appointments, the systemic factors underpinning the patient's trust in his regular GP should equivalently underpin his systemic trust in the substitute. Both physicians underwent standard training, are employed by the same medical practice, are held to the same ethical standards, are influenced by the same professional norms, and so on.[1]

However, this is not to say that both physicians are indeed equivalently trustworthy such that the patient would do equally well to rely on physician A as physician B. Idiosyncratic factors – individual expert characteristics, motivations, intentions, discrepancies in skill, etc. – make experts differently trustworthy in terms of idiosyncratic trust, and a novice may become familiar with more of these qualities over time. In this way, trust in an expert can be grounded in both idiosyncratic and systemic factors; a patient might trust a physician because the physician is embedded in the medical profession (systemic factor) and because in the past the patient has experienced the physician to be a thorough investigator and attentive to the patient's concerns (idiosyncratic factors). So, to summarize, all experts who occupy the same node in a network of systemic factors are interchangeable in terms of systemic trust while idiosyncratic factors can help to differentiate between experts above and beyond that threshold.

[1] Inversely, when an individual cannot be switched out for another, it is an indication that the system is not functioning well as a foundation for systemic trust. For instance, that it is possible for an elected US president to be a volatile leader says the US electoral system and the checks and balances in place to regulate executive powers do not underpin a high degree of systemic trust in US presidents. This is not to say that if the system were well-functioning, all individuals who succeed in being elected president will be equally trustworthy. Rather, a private citizen should be able to assume some minimum degree of trustworthiness. In other words, you may not agree with the elected leader on many points, but, in general, you can reasonably assume that the new leader will endeavor to serve the interests of the country and its citizens and will not easily be able to take steps to intentionally cause damage.

2. SYSTEMIC TRUST OVER IDIOSYNCRATIC TRUST

Overall, systemic trust is rooted in one's assumption that the network of systemic factors in which an expert is embedded is functioning well such that it has a positive regulating influence on the expert's performance. By positive regulating influence I mean that systemic factors serve to incentivize, mandate, restrict, encourage, or otherwise influence expert training and behavior in such a way that experts meet minimum standards for knowledge, skill, and ethical conduct in their given field (see chapter 6 for a more detailed discussion of well-functioning systems). In this section I argue that novice trust in experts is best understood within the framework of systemic trust instead of idiosyncratic trust, and that systemic trust should be the primary focus of those wishing to understand, analyze, and/or intervene in the dynamics of novice trust in expertise.

A quick reminder, over the next few pages I discuss systemic and idiosyncratic trust only with respect to human experts in order to avoid overcomplicating the conversation. I will bring the conversation back around to systemic trust in AI in section 3.

Systemic trust is important for those wishing to study novice trust in expertise for a couple of reasons. The first is a descriptive point: novices tend to trust experts based on assumptions about systemic factors more so than idiosyncratic factors, and therefore literature on trust in expertise should attend to systemic trust. Consider, for example, the basis of your own trust in the virologist who insists that a vaccine is safe, the climate expert who warns of the impacts of melting ice caps, or the astrophysicist who proclaims existence of black holes. You likely believe the claims of one or all of these experts while having very little reason to trust the individual speaker based on her idiosyncratic features. More so, seeking further evidence of the individual experts' training experience or intentions is not high on your priority list. Rather, you trust the virologist, the climate scientist, and/or the astrophysicist because you believe that they are each practitioners of science, and science is supposedly characterized by methodological norms that grant the outputs of science high epistemological status.[2]

Inversely, novice distrust in experts is often rooted in misgivings about the systemic factors and institutional frameworks in which the individual experts are embedded, not the experts themselves. For example, Maya Goldenberg (2021: 115) claims that widespread vaccine hesitancy is best explained by broad distrust of medical and scientific institutions. It is not that vaccine hesitant parents are stubborn or ignorant or reject the notion that an expert (like a physician) is more likely to be right about some things due to their specialized education, training, and experience. Rather these parents are concerned about medical research priorities (e.g. protecting population health as opposed to individual patient health) and about how financial incentives are able to influence medical research. If distrust in experts is rooted in doubts about the systemic factors that influence expert behavior and

[2] This assumption does, of course, enjoy its complications; I will touch on issues related to the supposed epistemic privilege of science throughout this section.

performance, then the proper response to improving novice trust in experts is to improve the social epistemic systems (or as Bouchard argues, improve the institutions) in which those experts are embedded.

Furthermore, it is systemic trust, not idiosyncratic trust, that is responsible for maintaining stable relationships of trust between novices and experts which, in turn, enables large-scale cooperation, teamwork, and social cohesion (Misztal, 1996; Stanley, 2003). As sociologist Georg Simmel (1978 [1900]: 191) explains, "without general trust that people have in each other, society itself would disintegrate." Yet, Simmel continues, very few relationships are "based upon what is known with certainty about another person, and very few relationships would endure if trust were not as strong as, or stronger than, rational proof or personal observation." Idiosyncratic trust can explain a person's willingness to rely on a small handful of people to make decisions, delegate tasks, and ask for advice, but there is a limit to the number of people a person can have strong reason to trust individually. Idiosyncratic trust requires considerable research and/or a high degree of familiarity with a trustee, and it is unlikely that a novice – or someone in whom a novice does hold strong individual trust, like a close friend or family member – will be able to meet these requirements for every expert the novice seeks to consult. Therefore, large-scale collaboration and teamwork requires novices to have reason to trust others without referencing idiosyncratic characteristics. General systemic trust, not idiosyncratic trust, is the very fabric of society.

Literature on novice trust in expertise and expert testimony does not, however, seem to reflect the centrality of systemic trust to novice trust in experts. Rather, idiosyncratic trust – trust based in analyzing the quality of expert characteristics – takes center stage as epistemologists discuss heuristics for novice appraisal of expert trustworthiness (*inter alia* see Walton, 1997; Goldman, 2001; Matheson, 2005; Anderson, 2011; Scholz, 2009; Martini, 2014; Lane, 2014). For example, in one highly influential paper, Elizabeth Anderson (2011) argues that a novice can evaluate an expert's trustworthiness by assessing the expert's expertise (knowledge and skill), honesty, and epistemic responsibility via a set of publicly accessible criteria. By publicly accessible criteria, Anderson means criteria that can be investigated and applied by people of ordinary education (no more than a high school education) who have access to the Web. In what follows I work through Anderson's criteria to demonstrate why the present focus on idiosyncratic trust is disproportionate to its importance, while, on the other hand, systemic trust is overlooked.

First, to appraise levels of expertise, Anderson says that novices can turn to expert credentials and certification and can infer the extent to which an expert is a leader in her field by examining expert biographical and bibliographic information. All this information can be found on the Web. Note, however, that as indicators of expertise, credentials and certifications derive their value from the systemic factors that back them (e.g. the educational standards enforced by certifying institutions or the methods employed to examine putative experts), yet Anderson does not extend her discussion into systemic backing. I will return to this point shortly.

Second, to appraise expert honesty, Anderson suggests novices look for evidence of the following factors that would discredit the expert:

(a) Conflicts of interest, such as receiving funds from agents who have a stake in getting people to believe a particular claim.
(b) Evidence of previous scientific dishonesty, such as plagiarism, faking experiments or data, and repeatedly citing research that does not support one's claims.
(c) Evidence of misleading statements, such as cherry-picking data or other misleading use of statistics, or taking quotations out of context.
(d) Persistently misrepresenting the arguments and claims of scientific opponents, or making false accusations of dishonesty against them (Anderson, 2011: 147).

Anderson does not expand much on how novices should investigate these criteria. She acknowledges that "some cases of dishonesty are difficult for laypersons to assess" but otherwise states that "other [cases], in which the evidence is readily available through the Web and verifiable without specialized knowledge, are clearly accessible to laypersons" (147).

Finally, to evaluate an expert's epistemic responsibility, Anderson explains that laypersons are capable of investigating whether experts are guilty of any of the following:

(a) Evasion of peer-review: refusing to share data for no good reason; refusing to reveal one's methods and procedures in enough detail to permit others' replication of one's experiments; failing to submit research to peer-reviewed journals; publicizing one's ideas in the press or in political circles before making one's case before experts.
(b) Dialogic irrationality: continuing to repeat claims after they have been publicly refuted by others, without responding to the refutations.
(c) Advancing crackpot theories in domains other than the one under investigation – for example, that HIV does not cause AIDS.
(d) Voluntarily associating with crackpots – e.g., publishing their work, or placing one's own work for publication in their venues (Anderson, 2011: 147-8).

Again, Anderson explains that each criterion can be investigated on the Web, and Anderson provides examples of how novices can evaluate exchanges between putative experts for evidence of dialogic irrationality without being familiar with the topic of conversation.

I grant that Anderson demonstrates that it is within the average adult's cognitive ability to mount investigations into an expert's idiosyncratic features. However, such investigations require a non-trivial amount of effort – more than lay people should be expected to exert. As I have argued elsewhere, being well-informed is a privilege of time and attention that most people do not have at their disposal (Seger et. al., 2020). Accordingly, day to day most novices do not spend significant time searching for and evaluating evidence to ground individual trust in each expert they encounter. Rather, the typical novice relies on an expert being one of a type, a skilled and knowledgeable individual embedded in a network of systemic factors that has regulatory influence over the expert's behavior ensuring the expert maintains an acceptably high level of knowledge and skill in her field and exercises her expertise responsibly in service of others.

However, it would be uncharitable to assume that Anderson expects that all or even most instances of novice trust in experts are based on the kinds of individual research she outlines. A better

interpretation of her work is perhaps as a study of idiosyncratic trust for when systemic trust does not provide an adequate foundation for deciding when to trust an expert. This is the case, for instance, in Alvin Goldman's (2001) Novice/2-expert challenge. Goldman asks how a novice should decide which expert to trust when two putative experts in the same field present competing claims. Systemic trust is of little use if the two experts seem to occupy the same node in a network of systemic factors. When experts are interchangeable in terms of systemic trust, idiosyncratic trust becomes an important differentiator. Similarly, Anderson's work can be interpreted as suggesting methods for ensuring that the everyday systemic trust novices place in experts is well-placed, especially in the case that trust is challenged. In this way, discussions of foundations of idiosyncratic trust as put forth by Goldman and Anderson are compatible with and complementary to my focus on systemic trust.

This is all fine. My aim is not to say that idiosyncratic trust is unimportant. Especially in cases where systemic trust cannot differentiate between conflicting experts or where an expert's trustworthiness is challenged, investigations into publicly accessible foundations for individual trust in experts is certainly a valuable avenue of research. Rather, my goal is to show that (a) the majority of literature on trust in expertise which focuses on the foundations of idiosyncratic trust is relevant only to a minority of cases, and to point out that (b) many idiosyncratic factors are closely linked to systemic factors anyway – i.e. reasoning about idiosyncratic trust involves background reasoning and/or assumptions about factors influencing systemic trust.

With respect to (a), Anderson, Goldman, and others (*inter alia* see Walton, 1997; Matheson, 2005; Scholz, 2009; Martini, 2014; Lane, 2014) seem to confuse the normal case – that a novice's trust in an expert is underpinned by assumptions of systemic trust – with the unusual case – that novices feel the need to investigate individual factors in deciding to trust an individual expert. If systemic trust accounts for most cases of novice trust in experts, then philosophical emphasis on idiosyncratic trust misrepresents most novice-expert relationships.

Of course, it could be argued that while systemic trust may be foundational to how novices *do* trust experts, this is an unfortunate state-of-affairs because systemic trust is epistemically inferior to idiosyncratic trust; it is not how novices *should* trust experts. For instance, I already noted that idiosyncratic trust goes above and beyond systemic trust in differentiating between experts situated at the same node in a network of systemic factors. Another way to interpret this is that systemic features are not specific to the individual and therefore are a more precarious foundation for trust in any given individual than idiosyncratic features. Furthermore, networks of systemic factors do not always function well such that they render trust in all embedded experts epistemically well-placed. For example, some professional groups have been accused of operating like "old boys' clubs" in which admittance depends heavily on social capital, and, once admitted, members protect their own. (Cullen & Perez-Truglia, 2020). Such a dynamic paves the way for individuals of mediocre skill and questionable ethics to persist in their practice. As such, systemic trust may be descriptive of how novices primarily do trust, but not descriptive of how novices should trust. I propose, however, that

such challenges to systemic trust as a basis for how novices can rationally decide to trust experts are challenges to specific cases of systemic trust. These are cases in which systemic factors are not in place that effectively influence speaker performance or in which the presence of such factors are not communicated to an audience so as to underpin internal epistemic justification for belief in speaker claims. Specific problematic cases can and should be addressed, and further analyses of what constitutes well-functioning systems will be discussed in Chapter 5.

Though systemic trust can be flawed, it is still of central importance to novice trust in experts. As discussed earlier, systemic trust based on assumptions about systemic factors and influences on individual behavior is essential to the cohesion and functioning of society. Therefore, the topic cannot be wholly set aside. But more importantly, point (b): many foundations for idiosyncratic trust quickly fall flat if not framed against the broader picture of systemic trust. Consider Anderson's strategies for lay appraisal of experts that are listed a couple pages back, or for the sake of simplicity, the five sources of evidence that Goldman (2001) outlines that a novice can use to appraise an expert. Anderson's criteria can be sorted into Goldman's more general categories:

(A) Arguments and explanations presented by experts in support of their claims or in critique of other experts' claims
(B) Evidence of experts' interests, and biases
(C) Agreement from additional experts
(D) Meta-expert appraisal (reflected in formal credentials granted by meta-expert institutions)
(E) Evidence of experts' past "track records" (Goldman, 2001: 93).

First, with regard to (A), as I explained in Chapter 2, explanations and arguments presented by experts are necessarily of limited justificatory value to novices. That is, due to the greater knowledge and superior skill and/or reasoning abilities of experts relative to novices, explanations experts offer for their claims that are cognitively accessible to novices will necessarily be limited in their (the explanation's) ability to help novices justify their belief in an expert's claims. By extension, the explanations will also be limited in their ability to help novices decide which experts are epistemically trustworthy. A novice's confidence in either the veracity of an expert's claim or in the premises of the expert's explanation thereof will be based in the novice's belief that the expert is one of a type – an individual whose claims to knowledge and skill are maintained by the standards and requirements of her profession and employer – rather than the novice's own appraisal of the expert's assertions.

There is a similar challenge with regard to (B). First, as Anderson acknowledges, it can be difficult for novices to appraise indications of expert interests and biases such as the use of misleading statistics, cherry picking data, or misrepresenting other experts' arguments. Second, it can be difficult to determine whether indications of an expert's interests and biases are also indicative of an expert's trustworthiness as a source of information. For example, Anderson proposes that novices can look into an expert's "conflicts of interest, such as receiving funds from agents who have a stake in getting people to believe a particular claim" (147). But it is not entirely clear if and how something like

receiving funds from an interested party actually influences the epistemic value of an expert's claims. For instance, that a cancer research lab receives funding from a tobacco company to research cancer treatments does not necessarily mean the research coming out of the lab is flawed or that experts in that lab are less epistemically trustworthy. If the researchers follow scientific standards of peer review, accepted scientific process, etc., then their outputs about cancer treatment are sound according to the scientific standards. One could argue that the tobacco funding may have driven the lab to research lung cancer treatment instead of prevention (and perhaps the latter, would have been preferable). In such a case, the choice of what topic to research in the first place may have been influenced by the researchers' adherence to the funders' wishes, but the science itself could still be epistemically sound, and the individual experts still trustworthy sources of information regarding recent advances in lung cancer treatment. Such a line of reasoning may not be so clear to a novice who is struck by the more obvious antagonism between lung health and smoking tobacco products. Accordingly, the ability of (B) to help ground idiosyncratic trust in an expert also assumes systemic factors are in place to positively influence expert behavior. Expert interests and biases are influenced by extrinsic systemic factors like financial incentives, and intrinsic systemic factors like professional community norms and expectations. Furthermore, any institutional mechanisms in place to handle conflicts of interest and/or mitigate their impact are, of course, systemic features. So, drawing conclusions about individual expert interests and motivations will often build on a base level of systemic trust and involve reasoning about the systems in which the expert is embedded.

(C), (D) and (E) similarly build on systemic trust. For example (C) is about corroboration by additional experts. The general idea Goldman puts forth is that the greater number of experts who agree with a putative expert having come to the same conclusion by at least partially independent means, the more likely it is that the putative expert's claims are true, or at least well-founded, according to the standards of the field, and therefore, the more justified the novice is in trusting that putative expert. But note that Goldman is still talking about the agreement of additional *experts*, not the agreement of just anyone. The agreement of additional experts is of greater value than the agreement of members of the general public because it is assumed that the beliefs of experts will be guided by systemic factors like higher epistemic standards and/or more rigorous training in the area of expertise.

I have already discussed (D): expert credentials and certifications are only meaningful if one assumes the well-functioning of the networks of systemic factors that back those credentials. It is worth adding that cases in which novices look to expert credentials and certifications as evidence of an expert's epistemic trustworthiness are often discussed under the heading of "meta-expert appraisal" coined by Goldman (2001; see Collins & Evans, 2007; Collins & Weinel, 2011; Goodwin, 2011; Kutrovatz & Zemplen, 2011). Meta-experts are experts at judging the expertise of others (their knowledge, skill, biases, etc.). The idea is that where a novice is not able to judge the epistemic trustworthiness of an expert for herself, she may instead turn to the evaluations – summarized in the

form of credentials and certifications – of meta-experts who are supposedly better suited to the task. As evidence of expert trustworthiness, meta-expert appraisal is limited, however, in that meta-experts are yet another class of expert that a novice must decide whether to believe. We are faced with a challenge of infinite regress: how can novices epistemically justify their belief in meta-experts? They would have to seek the advice of meta-meta-experts, and then meta-meta-meta-experts, and then meta-meta-meta-meta-experts, and so on. But this is absurd. If credentials and certifications provide meaningful evidence of epistemic trustworthiness, it is not because there exist infinite chains of meta-experts attesting to them, but because finite groups of meta-experts can be organized into credential-granting meta-expert institutions that are trustworthy evaluators of expertise in and of themselves. The capacity of any meta-expert institution (like an educational institution or professional organization) to reliably identify epistemically trustworthy experts and grant credentials and certifications accordingly is not reducible to the trustworthiness of its constituent members. Rather, it is resultant from the expertise of the constituent members as well as the organizational structures that dictate how those members interact, the measures and procedures they use to evaluate putative experts, and the consensus forming procedures they employ to coordinate and, where necessary, reconcile opinions.[3] Any certifications and credentials conferred by a meta-expert institution therefore testify to the group's decision as independent from that of any individual, and as such, trust in expertise via meta-expert appraisal is inextricably intertwined with systemic trust in meta-expert institutions.[4]

Finally, there is (E), evidence of past track record. Of all the sources of evidence a novice could draw upon to appraise expert trustworthiness, it would seem the track record would rely least, if at all, on an expert's systemic backing. However, as discussed in the introductory chapter, it is important to consider where novices get their information regarding track record. It could partially be from experience; a novice might have had good experiences with the expert in the past, so she trusts the expert to perform similarly in the future. However, a novice's personal experiences with an expert are a small sample size of the expert's activity, and such experiences would be completely lacking in first-time or one-off encounters. Evidence of past track record more likely comes in the form of an expert's reputation recognized and indicated through awards, promotions, credentials, and certificates. So again, when we take into account how expert track record is communicated to novices, novice trust in experts based on indications of track record presumes trust in the recognizing and reporting institutions.

[3] For more on group cognition and collective epistemology see Bird (2014), Lackey (2018) and Woolley et al. (2015). On consensus formation processes and judgment aggregation see Cariani (2011), Miller (2013), and Dietrich & Spiekermann (2013).

[4] Jennifer Lackey (2018) argues that a group itself can testify. Lackey argues against what she calls deflationary accounts of group testimony in which a group's ability to make an assertion is reduced to individual testimony as when an individual – a spokesperson – testifies on a group's behalf (Ludwig, 2014) or when a group testifies because its members are jointly committed to a given assertion (Gilbert, 2000, 2004; Fricker, 2012). For other accounts of group testimony that, like Lackey's, understand group testimony as distinct from individual testimony see Tollefsen (2007) and Hawley (2017).

I end this section with a brief analogy to help in understanding how systemic trust and idiosyncratic trust relate. Think of idiosyncratic trust like icing on a cake. The icing elevates the dessert, but icing would not be a good dessert on its own (at least in my opinion). Icing is most valuable when it builds upon a cake. Similarly, idiosyncratic trust elevates the level of trust in an expert by building on a foundation of systemic trust. Some idiosyncratic factors like expert credentials and certifications are only meaningful if they assume the well-functioning of networks of systemic factors, and, given the limited ability of novices to appraise individual expertise, idiosyncratic trust without a foundation of systemic trust is a precarious basis for novice trust in an expert. Therefore, discussions of novice trust in expertise are best framed in terms of systemic trust, with idiosyncratic trust providing the icing on the cake, and not the other way around.

3. SYSTEMIC TRUST IN AI

In the previous section I described systemic trust and argued that novice trust in experts is best described in terms of systemic trust. In this section I make a parallel case with respect to user trust in AI. Like trust in human experts, a person can trust an AI system based on idiosyncratic factors (e.g. track record or specific design features), and/or systemic factors (e.g. design standards or policy). I briefly argue that systemic trust should be considered the primary basis for user trust in AI as it is for novice trust in human experts (section 3.1), and I describe in more detail how systemic factors influence AI system trustworthiness (section 3.2).

I should note that there is much debate as to whether trust and trustworthiness are appropriate terms to use with respect to AI. I do not find this application of trust terminology problematic and will proceed in its use, though for those interested, I will take the space to justify my use of trust terminology in the next chapter.

3.1 User trust in AI is primarily systemic trust

As in the literature on novice trust in experts, most discussions about user trust in AI assume a model of individual trust in which users should trust AI systems that have certain characteristics more than AI systems that lack those characteristics. Among other things, an epistemically trustworthy AI system will be transparent or explainable (Ribiero et al., 2016), will be unbiased and fair (Benjamin, 2018), and have a track record of good past performance (Bishop & Trout, 2005).[5]

However, as when novices seek advice or insights from human experts, AI users are unlikely to be aware of specific details pertaining to the trustworthiness of the AI system. Mounting investigations into such details again requires the expenditure of a non-trivial amount of time and effort, and while some relevant information about idiosyncratic characteristics of an AI system may,

[5] On these and other idiosyncratic characteristics expected of trustworthy AI, also see the European Commission's (2019) *Ethics Guidelines for Trustworthy AI*.

in theory, be made publicly accessible via web search, they will likely be more difficult to interpret and therefore less meaningful to adults of an ordinary education than analogous information about human experts. For instance, if a company claims to have trained an expert system on diverse data sets, it is not immediately clear why this is important if one is not familiar with how machine learning works. On the other hand, most people will have some sense of why it is generally good for human experts to have extensive training experience in a variety of possible scenarios. Or to provide another example, as discussed at length in Chapters 1 and 2, AI explainability (an idiosyncratic feature of an AI system) is very limited in its ability to help lay users independently determine an AI system's likelihood of yielding true or epistemically well-founded outputs. A detailed explanation that is reflective of how the AI system derived its conclusion is most likely beyond what a typical user can comprehend let alone use to independently appraise the veracity of the AI system's outputs. On the other hand, a heavily simplified explanation will lack information for the user to appraise. Going forward, if an AI system's "being explainable" helps a user to justify her belief in AI system outputs, I posit that it is often because calling an AI system "explainable" functions like a credential. It suggests to the user that someone else has been able to appraise how the system derives its conclusions and has deemed that process to be epistemically safe and/or sensitive. Whether one ought to place any stock in an AI system being classified as "explainable" will, of course, depend on the label's systemic backing: how is explainability defined, how is it appraised, what is it appraised for, etc.

So overall, as AI applications become more commonplace, it should not be expected that trust in AI be grounded in idiosyncratic trust – trust based on the characteristics and features of individual AI applications. Rather, as with novice trust in human experts, most cases of user trust in AI will be cases of systemic trust – trust based on assumptions about systemic factors that influence how AI systems are developed and employed.

3.2 Foundations of systemic trust in AI

The more difficult question about systemic trust in AI is about how systemic trust – trust grounded in the intrinsic and extrinsic systemic factors that influence individual performance – relates to trust in AI.

Like systemic trust in human experts, systemic trust in AI is based on assumptions about the well-functioning of the systems by which the AI systems are produced and in which the AI systems are employed and managed. This similarity is made clear by distinguishing between the two subsystems in which systemic factors have an effect: *systems of production* (training, education, etc.) and *systems of employment*. In the human case there is often little distinction between systems of production and systems of employment. For example, a physician is both a product of the medical community and a member in its employ. As a product of the medical community the physician underwent training overseen by established physicians and was evaluated for the satisfaction of

specific standards of expert knowledge before being approved to practice medicine independently. As a member of the medical community, the physician continues to maintain the standards of her education, keeps up to date with evolving professional standards, and continues to adhere to the medical community's norms of conduct. If a person were to complete her medical training but opt to work as a ski instructor thereafter, she would not be employed by the medical community, but she would still retain her status as a product of the medical community. As such, anyone who were to query the ski instructor for her medical opinion might have some reason to accept the instructor's advice given her status as a product of the medical community but not in any capacity as a practicing member. So, while there is often significant overlap in an expert's relationship to an expert community as a product and/or member, these relationships in principle provide distinct reasons to accept the individual expert's claims.

In the case of AI, we can similarly look to systems of production and employment to help identify the factors influencing systemic trust, though the system in which an AI operates (its system of employment) will often be different than its system of production. For instance, a medical diagnostic system might be employed in a medical context but have been produced against the backdrop of the Silicon Valley tech mogul scene. I will discuss systemic trust in AI in the contexts of systems of production and systems of employment in turn.

3.2.1 Systems of production

Andrew Selbst and Solon Barocas (2018: 1130) note that "when we seek to evaluate the justifications for decision-making that rely on a machine learning model, we are really asking about the institutional and subjective process behind its development." The relationship between systems of AI development and production and systemic trust in AI is rather straightforward. If AI producers are subject to extrinsic factors – e.g. policies, incentives, and audits – and are susceptible to intrinsic influences – e.g. professional norms and values – then these systemic factors will have some impact on the trustworthiness of their AI products. Such systemic factors are, of course, defeasible foundations of trust – a well-meaning and tightly regulated group of AI developers could still produce inept algorithms – but this is the case with any "marker" of trustworthiness.

Systemic trust rooted in systems of AI production is a kind of transitive trust, or to use Philip Nickel's (2021: 496) terminology, it "is a linear trust relationship (USER→TECHNOLOGY →ENGINEER), but an indirect one". As Nickel explains, trust in a technological instrument can be understood as trust in the engineers, and that trust in the engineers is further derived from the institutions, policymakers, regulations etc. influencing the quality of the engineers' outputs (see also Nickel, Franssen & Kroes, 2010). Systemic trust rooted in systems of production is what allows an aspiring car owner to buy a new car and, while knowing very little about cars, to be confident that it is a roadworthy vehicle that meets legal safety standards. In the same way, systemic trust rooted in

systems of production could, in theory, ground a user's trust in an AI product. What those systemic factors should look like and whether they are in place are separate questions to be discussed in Chapter 5.

Of course, the interesting thing about AI – the thing that sparks more debate around the concept of trustworthy AI than trustworthy cars or coffee machines – is that AI-enabled technologies are specifically designed to function autonomously; they use information gathered from their environment to derive conclusions and to improve upon those derivation procedures in a process called machine learning. The end product is a technology that produces outputs for which no human engineer (or group of engineers) is responsible, and for which any chain of responsibility leading back from the system's outputs to its human developers is tenuous at best.[6] As such, it is not clear that the kind of linear trust relationship Nickel describes applies well to AI.

Recall, however, that Nickel has already given us a way around his conundrum in his account of norm-based entitlement (Nickel, 2013). Assurance need not be given, nor responsibility taken, for one to gain justification for believing a claim, or in the present case, an AI output. Rather, justification for belief can be rooted in the publicly visible norms of conduct that govern speaker behavior, and I have expanded on this idea to include the influence of other extrinsic systemic factors like educational rules and requirements. In the context of a human expert's production (education and training) an expert is taught skills and knowledge to enable her to think for herself, to draw conclusions and make decisions autonomously. At no point would we argue, however, that trust in the expert is a direct transmission of the trust we hold in the expert's past instructors, nor would we hold those instructors responsible for the expert's mistakes (nor give them full credit for the expert's successes). Regardless, it is uncontroversial that human expert systems of production (systems of education and training) have impact on the quality of service that the expert is able to provide and therefore have impact on the epistemic trustworthiness of the expert. Similarly, while AI systems do not respond to policy or internalize norms in the same way as human experts, there are still grounds for transitive trust from producer to AI system in the sense that intrinsic and extrinsic systemic factors that influence AI researcher and developer activities will have an indirect influence on the quality of the AI systems they produce. As with human experts, this transitive trust does not depend upon there being a teacher, trainer, or developer who can be held directly responsible for the AI's performance.

3.2.2 Systems of employment

The relationship between systems of employment and systemic trust is less obvious in the case of AI than in the case of human experts. Admittedly it is a bit of a stretch to say that an AI system is a member of an expert community or to argue that an AI system is individually responsive

[6] For more on complications to accountability and responsibility posed by autonomous AI systems see Vedder & Naudts (2017).

to extrinsic and intrinsic factors influencing its community of employment in the same way as a physician, for example. However, it is still the case, as Brent Mittelstadt (2019: 503) writes, that "AI inevitably becomes entangled in the ethical and political dimensions of vocations and practices in which it is embedded." I Identify two ways in which systemic trust pertains to AI at the level of employment.

The first way is by bootstrapping on human expert responsiveness to systemic factors.[7] One way expert systems can be built to mimic human expert performance is by training those systems on examples of human expert behavior. For instance, a medical treatment planning system might be trained on data regarding previous patient clinical information, treatments, and treatment outcomes to produce treatment recommendations for new patients. It is frequently pointed out in such cases an AI system's outputs should only be expected to be as good or bad, accurate or inaccurate, unbiased or biased as the human-sourced data on which it was trained. "Garbage in, garbage out," the saying goes. To expand on this observation, if an AI system is employed in an area that historically suffers from biased decision making by human experts, then having been trained on that historical data, the AI system may be expected to reflect those biases resulting in "algorithmic oppression" of historically marginalized people (Noble, 2018). For example, law enforcement in the United States is racially biased; African Americans are arrested more frequently, incarcerated for longer, and are less likely to be granted parole than white people having committed equivalent crimes. These racial biases were clearly reflected by the COMPAS recidivism software, an expert system used to predict the likelihood of criminal repeat offense and to provide parole recommendation for incarcerated criminals (Brennan et. al. 2009; Angwin et.al. 2016). Black defendants were twice as likely to be classified as at high risk of repeat offense than their white counterparts, and white defendants were mislabeled as low risk far more often than black defendants. COMPAS displayed racial biases because the American legal and law enforcement systems are racially biased. Similar cases of racial disparity also exist in medical systems which are perpetuated in the form of algorithmic bias and oppression. As Ruha Benjamin (2019) writes, "data used to train automated systems are typically historic and, in the context of health care, this history entails segregated hospital facilities, racist medical curricula, and unequal insurance structures, among other factors." (422). Benjamin notes that the upside, however, is that if extrinsic and intrinsic measures were implemented that effectively minimized these biases, the change would be reflected in the expert system's performance as well (See also Hampton, 2021). The idea here is that AI systems mimic the behavior of human experts in their domains of employment. If those

[7] Some clarification on the term 'bootstrapping' is needed. Here I use the term according to its colloquial definition: to get (oneself or something) into or out of a situation using existing resources. This is not to be confused with how the term "bootstrapping" is often used in AI research. In AI research bootstrapping refers either to a statistical process by which statistical analysis are made on estimates of statistical data sets (a kind of meta-analysis that help avoid overfitting of data), or bootstrapping refers to the more abstract possibility that AI system, having improved itself, will then be better at improving itself leading to exponential growth in intelligence.

behaviors are plagued by historical systemic biases, then so will be the conduct of the AI systems which, in turn, undermines the AI systems' epistemic trustworthiness.[8]

Second, systemic trust pertains to AI because a user might trust (or distrust) the community in which the AI system is employed; she trusts that the decision to employ the AI system to a particular role was well reasoned by someone (or a group of someones) in the same way as she trusts that the decision to employ certain medical staff was well-reasoned by someone or that the decision to employ certain medical supplies (scalpels, blood pressure monitors, surgical gloves, etc.) was well reasoned by someone. There is nothing importantly different between how systemic trust is grounded for human experts, AI systems, or simple instruments. Consider again the case of substituting GPs in a medical practice. A patient's trust in the replacement GP is grounded in her belief that the systems that regulate who is employed as a practicing physician ensure that the substitute meets minimum standards of expertise, ethical conduct, and patient care. But imagine now that instead of one GP being substituted for another, the GP is instead substituted for an AI system. For instance, the task of visually screening skin spots for potential cancer cases is delegated to a diagnostic AI. In the same way as a patient derives systemic trust in a substitute GP from her confidence in the employing medical community, so too may she derive systemic trust in the newly instated AI system. Assuming the medical systems and infrastructures are functioning well, when a patient walks into a surgery she may trust decisions about the employment of human experts and material instruments (from AI systems to scalpels) have been reasoned through so as to meet the profession's standards of conduct and care.

This does not mean, however, that systemic trust held in each of these entities is equivalent. As described in Section 1, entities are interchangeable in terms of systemic trust when they occupy the same node in a network of systemic factors. However, human experts and AI-enabled expert systems occupy different nodes in the networks of systemic factors that influence their performance. In the context of production, human experts and AI systems have different backgrounds of training and development and are regulated by different rules and policies. Furthermore, in the context of employment, strategies likely differ for vetting human experts for employment than for vetting various non-human instruments and tools, or an institution may simply be more familiar with evaluating human experts for employment than AI counterparts. As such, a medical patient, for example, may have more faith in a medical institution's ability to identify and employ quality human

[8] One might point out that the AI system biases I have described seem to have more to do with moral aspects of trust than epistemic aspects. Afterall, an AI system could sift through resumes by excluding candidates from the interview round based, among other things, on race and gender, yet in so doing still pick excellent candidates for the job. I am convinced, however, by Gürol Irzik and Faik Kurtulmus's (2018, 2019) notion of enhanced epistemic trust for which, the authors argue, the alignment of trustee values with those of the person of public being asked to grant their trust is key. In short, Irzik and Kurtulmus's argument stems from their analysis of inductive risk problems which, briefly put, point out that non-epistemic values play a role in determining the thresholds of evidence that must be met for the acceptance or rejection of hypotheses (Rudner, 1953). I discuss enhanced epistemic trust further in chapter 5.

medical experts than to identify quality AI tools.[9] Therefore, my point is not that human experts and AI-systems are interchangeable in terms of systemic trust – they are not – but (a) that systemic trust is the primary basis for novice/user trust in both human experts and AI, and (b) that systemic factors influence both human expert and AI system trustworthiness in their contexts of production and employment.

4. CONCLUSION

In this chapter I provided a general overview of what systemic trust is, how it works, and why it is an important area of investigation if we are to better understand the foundations of novice trust in expertise and user trust in AI; systemic trust, not individual trust, underpins most cases of novice and user trust in experts and AI respectively. In Chapter 5, I investigate further what makes a social epistemic system function well such that it attracts user and novice trust to entities embedded in that system and renders that trust well-placed. But before I proceed to Chapter 5, I present in Chapter 4 a brief interlude in which I address concerns regarding the use of trust terminology with respect to AI.

[9] Domenicucci and Holton (2017) differentiate between two-place trust (A trusts B) and three-place trust (A trusts B to X). Philosophical orthodoxy is that of three-place trust, the idea being that A may have very good reason to trust B (a dog loving alcoholic) with regard to X (feeding the dog) but not with regard to Y (not drinking). Two-place trust, on the other hand, is the kind of trust that is uniform with respect to whatever a trustee does. Domenicucci and Holton describe two-place trust as arising in cases of love and close friendship.

4

A Brief Interlude on Trust

In the previous chapter I explained that systemic trust plays a prominent role in novice-expert relationships, and I have argued it plays a similar role in user-AI trust relationships as well. Before I discuss the nature of epistemically well-functioning systems in Chapter 5, in this brief interlude, I address a few concerns regarding the use of trust terminology with respect to AI.

What I describe as systemic trust in AI may seem more like trust in AI producers, regulators, and employers than trust in AI itself. I do not see this as a problem. As far as I am concerned, if a person trusts the producer and/or employer, then she has reason to trust the product. For example, if a physician trusts Johnson & Johnson as a producer of reliable and safe medical devices, then she might also say "I trust medical instruments produced by Johnson & Johnson". That being said, many will still take issue with my use of trust terminology with reference to AI and instruments and prefer I say that if a person trusts the producer and/or employer, then she has reason to *rely* on the product. Trust, they would argue, is a term reserved for a specific class of entity that does not include artifacts like instruments of AI systems. It is this second complaint that I attend to in this chapter. If, however, the reader does not find my use of trust terminology with respect to AI problematic, she may wish to skip directly to Chapter 5. This interlude is for those who still need convincing.

In this chapter I first review existing philosophical perspectives on the trust-reliance distinction, and I make an inductive argument against the possibility of settling on a single notion of trust that conclusively includes or excludes AI systems (or any other artifact for that matter) as an object of trust. I then present my own theory on the distinction between trust and reliance. In so doing I also distinguish cases of reliance from instances of mere use. I show how this three-part distinction

accommodates prominent accounts of trust while avoiding notable pitfalls. It is a distinction based in the reasons people hold for offloading some activity, cognitive process, or responsibility to an external entity, not the kind of thing that external entity is.

Note that in this chapter I am intentionally lax about maintaining the distinction between epistemic trust and practical trust as laid out in Chapter 3. Where epistemic trust is rooted in reasons to believe a trustee will produce true or epistemically well-founded outputs, and practical trust is rooted in reasons to believe a trustee will behave or act as one would hope, here I generalize both practical trust and epistemic trust as being rooted in in the reasons one holds to believe a trustee will "perform" well. Performance may refer either to the trustee' behavior or to the epistemic quality of the trustee's claims. For the purpose of this chapter, I conflate the two types of trust because whether one is discussing epistemic trust or practical trust, the complaints against applying trust terminology to artifacts are largely the same. Where I make a point or provide an example that applies exclusively to one type of trust or the other, I make a note.

1. EXISTING VIEWS ON TRUST V. RELIANCE

The question of when and where trust is an appropriate attitude to hold is a well-trodden philosophical topic. The standard approach to attempting an answer is first to defend a definition of trust and then, according to that definition, to determine whether trust is an appropriate attitude to hold towards a specific entity. In this manner, debate unfolds as to whether individual humans are always proper targets of trust (Fricker, 2012; Nickel, 2012), and whether trust is an appropriate attitude to hold toward collective entities like companies or states (Ludwig, 2014; Lackey, 2014; Fricker, 2012; Gilbert, 2004), toward animals (Lewis & Marsh, 2022), or toward artifacts like AI systems and toaster ovens (Nickel et al., 2010; Nickel, 2011 Taddeo, 2009; Coeckelbergh, 2012, Lewis & Marsh, 2022; Bryson, 2018b).

As a most basic requirement, trust is a willingness to depend upon another party to perform in a certain way. This much is uncontroversial. However, it is an incomplete account of trust according to most. Even the least restrictive rational choice accounts of trust, such as James Coleman's (1994) cost-benefit account, understand trust as involving some kind of risk analysis. According to Coleman, for instance, trust is a willingness to rely on another party in order to serve one's needs, and one grants their trust after deciding that the potential gains of trusting outweigh the potential losses. An AI system would clearly qualify as a proper subject of trust under such an account.

However, most would argue that what Coleman describes is reliance, not trust. Indeed, much discussion in the trust literature hinges on this distinction, with most arguing that there is some epistemically or morally significant difference between the two. For example, one view of reliance, when held distinct from trust, is that reliance is a decision to offload decisions, tasks, etc. to an external actor based on evidence of the actor's past performance, while trust is understood as a

willingness to rely on an external actor in the absence of such evidence. As Midden and Huijts (2009: 744) write, “trust becomes the basis of decisions at the point when other assurances are not sufficiently available and (experience-based) confidence is lacking”. On this conception of trust, trustworthiness is an inferior basis on which to make decision about whether to rely than evidence of reliability; a decision to rely is based on substantial experiential evidence, but a decision to trust is to take a leap of faith where such evidence is lacking or too weak to merit reliance. For example, based on years of experience you could rely on your old general practitioner to provide accurate diagnoses and to answer your questions honestly; however, you have no option but to trust her new replacement, human or AI.

But there is something here that does not quite sit right. It also seems appropriate to say that the more evidence you have of an actor’s good past performance, the more willing you should be to trust the actor. The actor’s present and future performance is not guaranteed by past performance, but a good track record certainly provides reason to trust she will act as you hope. This intuitive observation points to a contrasting, and more widely employed, way of conceptualizing trust not as something inferior to reliance but as something that is greater than mere reliance.

Trust is almost always understood as reliance plus something extra, with that something extra providing some distinctively ethical reason for believing a trustee above and beyond evidence of mere reliability. For most philosophers of trust, the goal is to identify what that “something else” is. Suggestions include that trust requires opportunity for the trustee to betray the trustor (Holton, 1994; Hawley, 2017), or the possibility of the trustee being (in)sincere (O’Neill, 2007; Coady, 1992; Lackey, 2006), or harboring goodwill or ill will toward a trustor (Baier, 1994). The trouble is that attempts to be more specific about what differentiates trust from reliance quickly diverge and are easily subject to intuitive counterexample. I will briefly run through a few prominent examples and comment on why we are unlikely to ever settle on one, or on any combination thereof.

To begin, Annette Baier (1994) provides an influential account of trust that requires perceived goodwill on the part of the trustee (also see Karen Jones (1996)). Baier writes, “when I trust another, I depend on her goodwill toward me” (1994: 99). However, instances of trust void of goodwill are easy to find. Onora O’Neill (2002: 14) nicely counters Baier with the example of a patient who trusts a physician to approach the case professionally despite knowing that the physician finds her (the patient) annoying and bears her no goodwill. Furthermore, Richard Holton (1994: 65) notes that goodwill is not only unnecessary for trust, it is also insufficient; a conman may rely on his victim’s goodwill toward him to successfully execute a con, but it seems odd to say the conman therefore trusts his victim.

In a different but related vein, trust is considered appropriate when there is opportunity for feelings of betrayal and gratitude in response to how the trustee performs. Richard Holton (1994: 67) writes, "When you trust someone to do something, you rely on them to do it, and you regard that reliance in a certain way: you have a readiness to feel betrayal should it be disappointed, and gratitude

should it be upheld". Put another way, where there is no opportunity for betrayal or gratitude, one will only experience mere reliance toward the external entity – that is, reliance without trust.

Opportunity for betrayal and resentment may emerge where one believes a trustee holds goodwill towards them but this belief proves to be false. Alternatively, Baier (1986) suggests that feelings of gratitude and resentment are appropriate where some responsibility has been taken for a claim or action. For instance, with respect to trust in testimonial speakers, Katherine Hawley (2017: 273) specifies that "reliability or unreliability in the provision of (apparent) information becomes a matter of trustworthiness or untrustworthiness when the individual makes an assertion, or tells an audience something, thus somehow taking responsibility for what is said." The idea is that in taking responsibility one falls under some kind of moral obligation to deliver a certain kind or quality of performance. When one defaults on that obligation, those toward whom the responsibility was accepted have reason to feel slighted, betrayed or let down. Therefore, it would not make sense to feel betrayed, for instance, by a thermometer that provides an incorrect reading. Annoyed maybe, but not betrayed. There was no moral slight. The thermometer cannot assume responsibility for the accuracy of its readings and therefore has no moral obligations to perform well. Consequently, one might say that she relies on the thermometer to accurately measure the temperature, but an attribution of trust would be inappropriate.

Understanding trust as a response to opportunities for betrayal or gratitude stemming from accepted responsibility nicely circumnavigates challenges such as O'Neill's objection to goodwill accounts of trust. As part of their professional training and certification, physicians assume a responsibility toward their patients to provide a certain standard of medical care. So, even if the physician bears no goodwill toward a patient, it is still appropriate for the patient to trust the physician because there is opportunity for disappointment born from that betrayal of that responsibility.

But again, there are plenty of instances in which no responsibility is taken for a claim or action, but trust still seems the appropriate term to use. For example, imagine you found your grandmother's private journal in which she jotted down her thoughts while serving as a frontline nurse in World War II. In writing her private thoughts your grandmother assumed no responsibility toward anyone for the veracity of the journal's content with respect to the events of the War. Accordingly, you have no grounds for feeling betrayed if she ever did exaggerate her accounts or invent fictional stories. Yet, having known your grandmother to be a reflective person (the journaling type), and in the absence of any reason to think she would have misrepresented events in the records of her private thoughts, it still feels appropriate to say that you trust your grandmother to have provided an accurate account of life as a war nurse.[1]

[1] There are those who would argue that in the act of "speaking" one automatically takes responsibility for what they say (Pierce, 1932: 384; Williamson, 2000: 268-9). I maintain, however, that assuming responsibility requires that responsibility is held towards someone. That someone need not be specifically identified. For instance, if I were to write and publish a book about my experiences during the COVID-19 pandemic, then I assume some responsibility toward my present and future readers, whoever they may be, for accounts I provide

I have barely touched the surface of the vast literature dealing in different understandings of trust and the counterexamples that can be mounted against each. More so, each of the authors I have mentioned have, or I am sure could have, formulated responses to the counterexamples I have provided to which other philosophers have, or could have, responded. However, my aim is not to provide a comprehensive summary of the vast literature on trust, (for those interested, Judith Simon (2013) provides an excellent bibliographic guide) but to illustrate what Thomas Simpson (2012) has described as an inductive argument against a singular notion of trust. The examples above provided reason to believe that other attempts at a singular definition of trust will be similarly vulnerable to counterexample, and, as Simpson writes, "Counter-examples can be given so easily because there are so many ways the word may permissibly be used, and so it would be foolish to seek a single definition" (554). Simpson acknowledges that a singular definition of trust could be achieved by a disjunctive approach in which "you trust someone if and only if either you rely on their goodwill *or* you adopt the participant reactive stance towards them *or* you believe it is in their interests to be trustworthy *or* . . ." but he is quick to point out that by accommodating all intuitive notions of trust, a disjunctive definition ceases to provide any clarity and precision to philosophical debate (554).

Simpson's solution is to do away with the trust-reliance debate all together. Where a single, well-circumscribed notion of trust is lacking, Simpson recommends that we select a definition that applies to the discussion at hand from the array of distinctive forms of trust available. The important thing is stating what concept of trust we are using and why, not arguing that it is the one true concept. Simpson writes, "stipulating that this is the concept which is on the table for a particular discussion provides a starting point. Once this is done, there is no reason to be suspicious of an analysis of that particular form of trust... Clarity in philosophical debate requires specifying which type of trust is at issue" (Simpson, 2012: 566).

There is merit to Simpson's approach. Most of the time when we talk of trust, whether it is trust in other individuals, in groups, organizations, or states, in our pets, children, thermometers, or in AI, the notions of trust we have in mind will be somewhere in the milieu between the most unrestrictive views of trust, like Coleman's cost-benefit account (which easily accommodates non-human subjects of trust), and the most restrictive views which combine requirements for responsibility, autonomy, goodwill, opportunities for gratitude and betrayal, and so forth (and which do not easily accommodate non-human subjects of trust). Therefore, if we wish to discuss foundations of user trust in AI, our efforts are better spent finding where in the mix AI lies, and stipulating which notions of trust we will discuss, than by trying to articulate the one true notion of trust and reliance. For example, in the previous chapter I showed how the idea of trust in AI can be clarified by situating

---

and the claims I make. In the case of your grandmother, however, her words were intended to be kept private. Responsibility was explicitly assumed toward no one. Whether the journal still counts as an instance of speaking or testimony, without any responsibility for its content having been taken, is a separate question. (See Hawley, 2017; Lackey, 1999, 2008 Ch.2, and Nickel, 2012 for further discussion on the topic.)

AI systems within the networks of systemic factors influencing the contexts in which AI systems are produced and employed.

In line with Simpson's approach, I find the application of trust terminology to AI systems (or any other artifact for that matter) unproblematic so long as we are careful to stipulate what notion of trust we have in mind. However, I do not go so far as Simpson in doing away completely with the trust-reliance distinction. If trust terminology is appropriately applied to artifacts provided that the notion of trust in use is clearly laid out, then by the same logic, I might just as well be speaking in terms of reliance (so long as notion or reliance in use is clearly laid out). However, my decision to speak of AI systems in terms of trust instead of reliance is not merely a stylistic choice. I hold that speaking in terms of trust still communicates something subtle about the reasons behind one's willingness to believe or depend upon an external entity which is not communicated by speaking in terms of reliance.

2. TRUST, RELIANCE & USE

Current debate about whether trust can appropriately be held towards an AI system or other artifact hinges on whether the entity is endowed with necessary capabilities. For example, if trust requires goodwill, then the trustee must be capable of holding goodwill towards others. If trust requires opportunity for gratitude or resentment, then the trustee must be capable of taking responsibility for actions and claims. In this section however, I explain my use of trust terminology toward AI by putting forward an alternative account of trust as something distinct from reliance. However, instead of the trust-reliance distinction being based on the kinds of reasons one holds for her willingness to depend on an external entity, which is the norm, my distinction is based on the kinds of activities to which trust and reliance refer.

To trust, I argue, is to adopt a mental state toward an external entity in which one is willing to depend on that entity. This mental state is based in beliefs about how the entity is likely to perform, and those beliefs may be informed by a variety of factors such as perceived goodwill, responsibility, sincerity, assurances given, or evidence of good track-record. Following Simpson, my account of trust is flexible with respect to which factors inform the appraisal of any given trustee so long as they are clearly stated and shown to be relevant to the case at hand. To rely, on the other hand, involves actively depending on an external entity to perform some task or function. Reliance is independent from trust, I will argue, because one can choose to rely on an external entity without having an attitude of trust towards that entity. Inversely, one can hold the mental state of trust toward an external entity without engaging in an act of reliance.

I will also differentiate between reliance and a third term, *use*. I argue that all three terms, trust, reliance, and use, can be appropriately employed with respect to humans, simple tools, or anything in between depending on the reasons (or lack thereof) people have for choosing to believe or

depend upon an external entity. I begin first by distinguishing between use and reliance (Section 2.1), and then between reliance and trust (Section 2.2). Box 4.1 at the end of Section 2.2 provides a summary of terms.

2.1 Use v. Reliance

Traditionally, when trust is not considered an appropriate attitude to hold towards some external entity, the default, then, is to speak in terms of reliance. To rely is to depend. One can depend on their bicycle to transport them to work, on a weather simulation to accurately predict the week's weather, or on one's PhD supervisor to provide helpful feedback on chapter drafts.

While reliance can be used with respect to either humans or artifacts, I propose, however, that it is not always appropriate to do so. For example, I think it is odd to say that a person *relies* on their toaster to make toast every morning. They just *use* the toaster. Similarly, when I pick up a pen to jot down a note, I do not rely on the pen, I just *use* the pen. Reliance implies something about having considered the weight of the consequences of the object of reliance failing in its role. For example, imagine you show up to a 3-hour written exam with one pen. In this case I would say you are relying on that pen quite heavily. God forbid it sputters out! Use, on the other hand, is ambivalent to any such consideration. Use is the act of utilizing someone or something to achieve an end, and mere use (use without reliance) is an unreflective activity that does not involve the consideration of consequences or possibility of failure.

What was once an instance of reliance may become an instance of use if the chances of failure have become so small, or the consequences of failure so inconsequential, that they cease to invite consideration. Such might be the case, for instance, if a person acquires strong evidence of an instrument's reliability, where reliability is the characteristic of something or someone that consistently acts or performs the way you expect. For example, while working from home, Josh's reliance on his home Wi-Fi might turn to mere use after several months of glitch free Wi-Fi service. He simply stops considering the possibility of dropped conference calls.

Inversely, instances of mere use may become instances of reliance because one notices the object of use has become prone to failure. For example, if your toaster starts to occasionally burn your toast, then what was once a mindless morning task becomes a considered decision: Do I try to make toast and possibly waste bread, or do I make porridge, my only other breakfast option, which I detest?

Instruments and people alike can be both relied upon or used. A person who is used often is sometimes referred to as a 'tool'. It is generally considered rude to call someone a 'tool' because tools are typically inanimate objects people employ to help achieve some end, so to call someone a tool is to imply that they lack moral agency and autonomous decision-making capacities. While social norms dictate that it is rude to call someone a tool, I propose people are used like tools all the time and that doing so is most often just fine. For example, when I cook dinner with my eight-year-old niece and I

say, “Evelyn, pass me a spoon,” I do not stop to think about whether Evelyn could or would pass me the spoon. In this instance I am unapologetically using her as a tool, specifically, a spoon passer. But later, when I ask Evelyn to pass me a sharp knife, I consider the request carefully. Evelyn and I have talked about knife safety on several occasions, and she generally demonstrates sensible behavior when working with me in the kitchen. I also know Evelyn cares about me, so I believe she will take extra care to avoid inflicting injury. I am no longer using Evelyn as a tool because I am not merely using her. I have considered the risks and decided to *rely* on her to provide me with a knife. Over time my confidence in Evelyn may grow to a point that I stop entertaining the possibility of injury all together. At such a time, my reliance on Evelyn as a knife passer would again turn to use. In this way, using someone like a tool can be a compliment. It means they pose so little threat of failure or misconduct that you do not stop to consider the possibility.

2.2 Reliance v. Trust

Like reliance, I also consider trust to be the result of balanced consideration. The example of me asking Evelyn to pass the knife illustrates both an instance of reliance and an instance of trust (specifically practical trust), but it will be easier for me to describe the distinction between trust and reliance if I start with the ways in which trust and reliance come apart.

To begin, a person can rely on someone or something she believes is unreliable, but she cannot trust someone or something she believes is untrustworthy. This is because ‘to rely’ is an outward act – the act of depending on an external entity having considered the risks and benefits of doing so – while ‘to trust’ (inversely, ‘to distrust’) is a mental state describing one’s (un)willingness to rely on some external entity. More specifically, it is an internal attitude toward someone or something based on one’s beliefs about how the entity is likely to perform. For example, if you believe your toaster is liable to burn your toast then your internal attitude toward the toaster is one of distrust, but you may rely on the toaster nonetheless because it is the only one you have. Or to provide another example, a judge might distrust an AI recidivism prediction system if she is aware that the system demonstrates racial bias in its predictions, but she may decide to rely on it to direct her parole decisions anyway because doing so is considered standard practice in the field.

A person can also trust without the involvement of reliance. For example, I believe my physician is a competent practitioner of medicine who keeps up with the latest medical research, so I generally trust her to provide me with epistemically well-founded medical advice (this is an example of epistemic trust). However, unless I am currently seeking medical advice, or unless I seek medical advice frequently and consistently, I am not in the state of relying on my physician.

I posit that the distinction between trust and reliance can seem to blur because very often when we talk about someone choosing to rely on an external entity, it would be appropriate to talk about the person experiencing a mental state of trust as well. Such is the case when I ask Evelyn to

pass me a knife. I experience trust toward Evelyn because I believe things about her (her past performance, her sensibility, and goodwill) that make me think she is a trustworthy knife passer. My beliefs about Evelyn's trustworthiness also factor strongly into my decision to rely on her. It is only when a person chooses to rely on someone or something despite their internally held belief that doing so will likely end badly that it is appropriate to speak of reliance but not trust.

I propose that a second point of confusion between reliance and trust stems from an overlap between the associated adjectives 'reliability' and 'trustworthiness'. Reliability is the characteristic of someone or something that performs a particular function (epistemic or practical) consistently well, while trustworthiness is the characteristic of something that is deemed worthy of trust. One reason someone or something may be deemed worthy of trust is because she or it seems reliable, perhaps as indicated by certifications or as demonstrated by past performance. However, perceptions of trustworthiness may also be based on other factors such as beliefs about the trustee's interests and motivations, beliefs about her goodwill, or beliefs about the presence of systemic factors that influence performance quality.

Finally, it is interesting to note the difference in relationship between trust and trustworthiness and between reliance and reliability. As discussed, trust must be based in one's perception of trustworthiness because trust is a mental state, but one might choose to rely on something despite one's perception of its unreliability or untrustworthiness. The knock-on effect is that the degree of one's trust in a trustee is rooted in the perceived trustworthiness of the trustee. The more trustworthy a trustee is perceived to be by the trustor, the stronger the trustor's trusting attitude will be.

On the other hand, how heavily one relies on an external entity is not responsive to the perceived reliability of the entity. Rather, how heavily one relies hinges on two factors: first, the ease or difficulty with which one would be able complete the desired task if the object of reliance were to fail, and second, the severity of the consequences of failure. For example, that single pen I brought to my exam might be a terribly unreliable pen (so I do not trust it), but I rely on it quite heavily because I realize that without it I cannot complete the exam, and without completing the exam I would fail the course.[2] Alternatively, Evelyn may be a steadfastly reliable knife passer (so I trust her), but my reliance on her is minor because if she were to abandon me in the kitchen, I could retrieve a knife myself and keep cooking. Dinner would still be served.

[2] In several of my undergraduate courses there was a strict no-pen-borrowing policy once an exam had commenced. If a student did not bring adequate writing supplies – which were checked by exam moderators for cheat sheets rolled inside – the student would simply not be able to complete the exam.

Summary of Terms [Box 4.1]

Use: The act of utilizing someone or something to achieve an end. Mere use (use without reliance or trust) is an unreflective activity and does not involve the consideration of consequences or possibility of failure.

Reliance: The act of depending on someone or something to perform some task or function after considering potential consequences and possibility of failure.

Trust: A mental state describing one's willingness to rely upon a trustee without guarantee of how the trustee will perform. Specifically, the mental state is the internal attitude you have towards something/someone because of beliefs you hold about the actor (e.g. reliability, goodwill, aligned values, effectively motivated, systemic backing etc.) indicating it/she is likely to perform in the way you hope.

Reliability: The characteristic of something/someone that consistently performs the way you expect.

Trustworthiness: The characteristic of something that is deemed worthy of trust. You could say something is trustworthy because it is reliable, but reliability is just one reason someone might have to trust. Other factors may include indications of trustee goodwill, values, sincerity, motivations, systemic backing etc.

3. SOME CONSIDERATIONS

So far, I have presented a notion of trust as distinct from reliance and from use. Trust is an internal attitude held towards a trustee based on reasons one holds for believing a trustee is trustworthy. A person may or may not act on their mental state of trust by choosing to rely on the trustee. Put otherwise, reliance is an outward act of dependence which involves consideration of possible benefits or adverse consequences of dependence, and which may or may not be backed by an attitude of trust toward the object of dependence. Mere use, like reliance, involves dependence on some external entity, but unlike reliance, is void of consideration of failure or consequence.

My account of trust as distinct from reliance and from use accommodates, to my knowledge, all cognitivist accounts of trust. Cognitivists hold that trust is a mental state based in one's beliefs about trustee's trustworthiness (McMyler & Ogungbure, 2018). Within the cognitivist tradition I have presented a pluralist view of trust in which perceptions of trustworthiness may be influenced by a variety of factors ranging from trustee goodwill and responsibility to evidence of reliability or the efficacy of systemic factors on regulating trustee performance. Though, like Simpson, I emphasize that for clarity and precision of philosophical debate, the specific factors relevant to evaluating a trustee's trustworthiness must first be clearly articulated. It would not be appropriate, for example, to suggest that an AI system's trustworthiness be understood in terms of the system's goodwill toward the user. While goodwill accounts of trust – accounts in which goodwill toward a trustor is essential to the existence of trust – are compatible with my understanding of trust as a way of specifying the

general picture, goodwill accounts of trust do not apply in the specific case of AI trustees. However, as discussed in the previous chapter, trust in AI can be understood in terms of systemic trust – trust held in an individual because of beliefs about how networks of intrinsic and extrinsic systemic factors can influence or regulate individual behavior.

There is a point, however, at which I fail to be as happily pluralist about trust as Simpson. I acknowledge that my understanding of trust as a mental state responding to perceived trustworthiness conflicts with non-cognitivist accounts of trust. Non-cognitivists understand trust not as state of mind based on evidence that the trustee is likely to perform as one would hope, but as one's willingness to take a leap of faith to depend on someone or something in the absence of evidence of trustworthiness (McMyler & Ogungbure, 2018; Midden & Huijts, 2009). A decision to trust might instead be based in one's hope that granting trust will inspire desirable performance (Nickels 2012, 2017; Faulkner, 2007) or because one believes she has some duty to trust others (Holton, 1994), and the greater the leap of faith, the greater the instance of trust. For example, Karen Jones (2004: 5) presents an example in which a mother trusts her teenage daughter to look after the house for the weekend even though the daughter has failed to do so in the past and the mother has no evidence to suggest the daughter will do so this time either. Nonetheless, the mother trusts the daughter in hopes that doing so will ultimately inspire a behavioral shift. The idea is that great demonstrations of trust can engender trustworthiness. Richard Holton (1994: 63) presents a similar example in which a shop owner discovers that a new employee has been convicted of petty theft, yet the owner still trusts the employee to operate the till unsupervised. Holton writes, "and again it appears that you can [trust] without believing that he is trustworthy. Perhaps you think trust is the best way to draw him back into the moral community; perhaps you simply think it is the way you ought to treat one of your employees." The non-cognitivist idea that trust can be held where perceived trustworthiness is lacking (or indeed, where a person is perceived to be untrustworthy) directly conflicts with the view I have put forward that one's mental state of trust is proportional to one's perceived trustworthiness of the trustee. In fact, what non-cognitivists describe as trust – the willingness to depend in the absence of evidence of trustworthiness – is what I have described as mere reliance – reliance without trust. Mine is a unique point of view. Indeed, it is precisely such instances of "ethical trust" described above which people usually think of as being the furthest away from reliance; If one has no evidence to suggest the subject is reliable (or, indeed, if one has evidence to the contrary), it seems the only other option is to trust. I understand the non-cognitivist sentiment, I just think the terminology is confused.

I think some non-cognitivist accounts can be reconstrued as cognitive accounts. For example, in cases like Jones's and Holton's, to believe that someone will display more desirable behavior in response to being relied upon is to believe something about the person that indicates they will act as one would hope. In other words, beliefs about trust-responsiveness can be constitutive of one's perception of trustee trustworthiness.

I still find problematic, however, the implication that the greater the leap of faith – meaning the less supporting evidence there is to believe the source is trustworthy – the more trust that is involved in deciding to rely on a trustee. On this point I strictly maintain my independence from non-cognitivists. The feeling of insecurity one experiences when taking a large leap of faith – like asking the wayward teen to take care of the house for the weekend – is not a feeling associated with a high degree of trust, it is the feeling of precariously positioning oneself to rely heavily on someone or something without much reason to believe that it will go well. When the mother tells her wayward daughter "I trust you with the house", her statement is dishonest. The mother has no reason to believe her daughter is trustworthy (or only little reason, if we take trust-responsiveness as constitutive of trustworthiness). An honest statement would be "I'm relying on you" or, at best, "I trust you, but only a very little bit". Whether the mother ought to exaggerate the extent of her trust in order to elicit a desired response so that she may one day honestly say, "I trust you", is a separate question.

Where non-cognitivist and my cognitivist account of trust come apart, I opt for the cognitivist perspective for good reason. This dissertation is concerned with the internal epistemic justification a person can hold for believing the claim or output of an expert or AI system. If that internal justification for belief is to be based in epistemic trust, then it follows that one ought to hold internal reasons for that trust. Those reasons come in the form of beliefs about the expert/AI that suggest that she/it will perform as one would hope – by yielding true and/or epistemically well-founded outputs.

4. CONCLUSION: TRUST IN AI

As I have defined them, the terms trust, use, and reliance can all be used appropriately with reference to humans, AI systems, or other artifacts. It is appropriate to speak of reliance instead of mere use when deciding to depend on an external entity involves considering the consequences of the entity failing to perform its function. This is the case whether the external entity is human – like a physician being relied upon to provide medical advice – or artifact – like a medical AI system relied upon to recommend treatment plans. It is appropriate to speak of reliance without trust when a decision to depend is not based on any belief that the entity is trustworthy (that it or she would perform as one would hope), and one may speak of trust without reliance when one holds a mental state of trust toward the external entity but does not engage in active dependence on the external entity. Most often, however, it is appropriate to speak of trust and reliance simultaneously because decisions to rely are often based, at least in part, on one's perception of the trustee's trustworthiness.

In this dissertation I intentionally discuss trust in AI, as opposed to reliance, to be precise about why a person chooses to believe an AI system's outputs. I am interested in cases of reliance with trust, instances of AI employment in which the decision to rely on an AI system is because one believes something positive about the AI system that suggests one would do well to rely on the system, not because there is no other choice but to rely or because one does so without consideration.

Having positive reasons to underpin one's willingness to defer AI-enabled expert systems is key to epistemic security of user-AI interactions, epistemic security being the overarching conceptual concern of this dissertation. Epistemic security describes the state in which one is able to consistently identify epistemically well-founded information or epistemically trustworthy information sources.

5

Well-functioning Systems

This dissertation investigates how people can go about acquiring some internal epistemic justification for their beliefs in human expert claims and AI outputs. In Part 1 (Chapters 1 and 2) I first considered the role of explanations. I asked whether the explanations offered by experts and AI systems contain information that can help recipients appraise the veracity or epistemic well-foundedness of expert/system claims and outputs. My conclusion was overwhelmingly negative. The greater the epistemic divide between explainer and explainee, the lower the maximum potential justificatory value of an explanation. Expert-novice relationships are, by definition, epistemically imbalanced, and AI-lay user relationships follow suit. It must therefore be expected that expert-to-novice and AI-to-user explanations will most often be of low justificatory value. Accordingly, if it is possible for users and novices to acquire any significant internal epistemic justification for their beliefs in expert claims and AI output, that justification will have to be derived elsewhere.

In Part 2, I propose that epistemic justification can instead be primarily rooted in epistemic trust, more specifically, a particular species of epistemic trust which I call systemic trust. In Chapter 3, I provided a general overview of what systemic trust is, how it works, and why it is an important area of investigation if we are to better understand the foundations of novice trust in expertise and user trust in

AI. In brief, systemic trust is the trust held in a trustee not because of idiosyncratic considerations about the individual trustee's characteristics, intentions, or motivations, but because trustee performance is influenced by a larger system of interacting people, policies, and norms in the trustee's contexts of production and employment. In this way systemic trust applies to both human experts and AI.

I now return to the topic of systemic trust to address the remaining questions: What kinds of systems underpin systemic trust in individuals? What do these systems look like? How are they built? And finally, are any such systems in place to underpin systemic trust in AI? This final substantive chapter dives into the nitty-gritty of *epistemically well-functioning systems* (networks of systemic factors that effectively influence the activities of embedded entities so as to render epistemic trust in those entities well-placed) and explores how such systems function in the case of AI-enabled expert systems.

This chapter proceeds in four parts. In Section 1, I discuss how extrinsic and intrinsic systemic factors and the interactions therebetween influence individual human expert performance. I propose two premises which, when satisfied, define ideally well-functioning networks of systemic factors: the *sociological premise* and the *enhanced epistemological premise*. Building on this discussion of systemic trust in human experts, I then turn my attention to systemic trust in AI. In Sections 2 and 3, I evaluate the existing foundations for systemic trust in AI according to these two premises. Finally, in Section 4, I make recommendations for strengthening foundations for systemic trust in AI. Section 5 concludes.

Before I proceed, note that I have shifted from discussing the acquisition of *internal* epistemic justification for belief in AI and expert claims (the stated goal of this dissertation) to the conditions under which one's beliefs in AI and expert claims would be *externally* justified. The main motivation behind building epistemically well-functioning systems is to make the world such that when one decides to trust an individual – and for whatever reason she decides to trust that individual – that that trust is indeed epistemically well-placed. In other words, epistemically well-functioning systems make the act of trusting safe. To echo Greco's (2012) articulation of epistemic safety, one would be unlikely to go wrong in believing the claims of any expert embedded in an ideally epistemically well-functioning system.

Epistemic safety is a key component of epistemic security, the overarching conceptual concern of this dissertation. An ideally epistemically secure society is one in which individuals consistently consume true and/or epistemically well-founded information. There are two ways to work toward this goal. The first is by figuring out how to help people to sift out true statements from false statements or, where that is impracticable, to sift epistemically trustworthy sources of information from the epistemically untrustworthy. This first approach is about internal justification for belief; it is about helping individuals to be more sensitive to falsehoods, incompetence, and charlatans. The second approach to improving a society's epistemic security is to build networks of social epistemic factors that eliminate the need for information recipients to sift the wheat from the chaff. Rather, the aim is to make the world epistemically

safe such that novices and lay users may rest easy in the assumption that by virtue of being embedded in epistemically well-functioning systems, experts and AI systems can reasonably be relied upon.

The challenge is, of course, that neither strategy – working to improve the epistemic sensitivity of information recipients nor the epistemic safety of society – will ever be implemented to perfection. As such, the best plan for working towards an epistemically secure society is to make progress on both fronts. It is therefore worth setting the discussion of internal justification aside for a moment to discuss from an externalist perspective how various systemic factors interact to impact the epistemic trustworthiness of experts and AI systems. That said, because this dissertation set out with the goal of pursuing a better understanding of how people can acquire a degree of internal epistemic justification for their belief in AI and expert outputs, I will return to the topic of internal epistemic justification in the concluding section. In this chapter I argue that the best internal justification novices and lay users can acquire for their belief in expert claims and AI outputs will very often be their confidence in the epistemic safety of the systems of production and employment in which those experts and AI-enabled expert-systems are embedded.

1. FEATURES OF WELL-FUNCTIONING SYSTEMS

In Chapter 3 I argued that systemic trust is the primary way we ought to think of epistemic trust in both human experts and AI. Now I address the subsequent question – what kinds of systems are epistemically safe such that they render systemic trust in individuals well-placed? This is a question about how networks of various intrinsic and extrinsic systemic factors influence expert performance.

In recent years similar questions have been investigated by social epistemologists seeking to understand how different social structures, procedures, and institutions influence epistemic activities.[1] For example, what Alvin Goldman (2012) calls 'systems oriented social epistemology' specifically aims to "examine a system in question to see whether its mode of operation is genuinely conducive to the specified epistemic ends [and to] identify alternative organizational structures that might be epistemically superior to the existing systems." (Goldman, 2012: 229). For instance, Robert Merton (1973) notes that institutional science is driven by a priority rule in which scientists are rewarded (e.g. Nobel prizes) for being the first to make a discovery or advancement, and Philip Kitcher (1993) argues that such incentive structures that appeal to scientists' selfish motivations for priority and credit are more effective at helping

[1] Building on Rene Descartes (1637) and John Locke (1690), (Western) epistemology traditionally has a strong individualistic focus on evaluating how epistemic agents use their personal cognitive devices to investigate questions and seek truth without interacting with others. Social epistemology takes root in the mid-20th century with the Strong Program in sociology of science (Bruno Latour and Steve Woolgar, 1986) and the work of philosophers such as Richard Rorty (1979) on "social justification of belief" and Thomas Kuhn (1962) on social factors as necessary to settling disputes in science.

a scientific community to achieve its epistemic goals than incentive structures that primarily motivate scientists to attend to community needs. To provide another example, Helen Longino (2002: 129-134) proposes basic social epistemic factors that are necessary for a scientific community to be epistemically well-functioning such that its outputs count as objective. These factors include an open forum for critical discourse among diverse epistemic viewpoints, expected uptake of criticism by the community when it is warranted, respect for the tempered intellectual equality of those offering input (tempered with respect to their training and track records), and publicly available standards for evaluating input.

I find Stephen John's (2018) critique of systems oriented social epistemology to provide a useful entry point for thinking about the features of epistemically well-functioning systems. John points out that even if a community is set up to meet fixed epistemic standards, this does not tell us much about whether the outputs of that community should be believed. To make any such determination, the epistemic standards themselves must be scrutinized for whether they are truth conducive. One could imagine, for example, that a community of astrologers has mechanisms in place that effectively cultivate and enforce the epistemic standards of astrology within its membership (e.g. education and certification programs, incentive structures, and open forums for discussion.).[2] It is not guaranteed, however, that meeting these standards will yield outputs that should be believed. In contrast, we tend to believe the research outputs of astronomers not only because we assume that the institutions of science effectively enforce the epistemic standards of astronomy (and its subfields: mathematics, physics, and chemistry) within the astronomical research community, but because those epistemic standards have also been shown to be truth conducive.

Accordingly, John proposes a model of how non-experts learn (acquire knowledge) from scientists – experts in a scientific discipline – via a "two step inference" in which two premises must be satisfied. First, is the sociological premise: "Institutional structures are such that the best explanation for the factual content of some claim (made by a scientist, or group, or subject to some consensus) is that this claim meets scientific 'epistemic standards' for proper acceptance" (John 2018: 77). The gist of John's sociological premise is that the institutions of science can be structured in such a way that they ensure scientific claims made by individual experts meet certain epistemic standards set by those institutions.

John notes, however, that even if his sociological premise is satisfied it does not mean that a non-expert's belief in a scientist's claim is justified. It is possible that astrology is a very well-structured discipline such that when an astrologer makes a claim one may be confident that the claim meets the accepted 'epistemic standards' of astrology, but this does not mean that the claim should be believed. The epistemic standards of astrology must also be the right kind of standards, they must be conducive to truth.

[2] The International Society for Astrological Research (ISAR) holds annual conferences which are open to the public and has a certification program (CAPISAR) for "astrological proficiency" in collaboration with schools arounds the globe, and facilitates a "Global Director" reward program for astrologers who contribute to astrological research and publicising astrological methodologies. <www.isarastrology.org> (accessed March 3, 2022).

Accordingly, John puts forth a second premise, the epistemological premise: "If some claim meets scientific epistemic standards for proper acceptance, then I should accept that claim as well" (John 2018: 77).

I believe John's two-step inference is on the right track. An epistemically well-functioning system is not only one that effectively implements desired standards, but those standards have to be the right kind. They have to be truth conducive. More so, that a community is set up to effectively implement its epistemic standards does not say much about whether those standards are the right kind. I propose, however, that a fuller picture of what it means for a social epistemic system to be epistemically well-functioning can be achieved by combining John's two-step inference with Gürol Irzik and Faik Kurtulmus's (2018, 2019) notion of enhanced epistemic trust.

Irzik and Kurtulmus differentiate between basic and enhanced levels of epistemic trust novices can hold in scientists. The difference is related to evidence of scientist-novice value alignment. Basic epistemic trust in a scientist requires the scientist to have proper credentials which Irzik and Kurtulmus take as evidence of an expert's basic competence in an area of expertise. In line with John's "two step inference", I would add to Irzik and Kurtulmus's account that taking credentials and certifications as a foundation for basic trust requires the satisfaction of the sociological and epistemological premises; the certifying institutions and practices must be such that credentials are only conferred to experts who do indeed meet high thresholds of knowledge and skill (the sociological premise must be satisfied) and who, by exercising their expertise well, one would do well to rely upon, assuming one is interested in obtaining epistemically well-founded information (the epistemological premise must be satisfied too).

Enhanced epistemic trust goes beyond this modified account of basic trust. As Irzik and Kurtulmus describe it, enhanced epistemic trust not only requires that the trustee have proper credentials (and, I have added, that she utilizes the knowledge and skill those credentials represent well), but that there also be evidence that the trustee's values align with the values of the person or public being asked to grant their trust. In other words, enhanced epistemic trust is warranted where epistemic standards are not only truth conducive but are also responsive to the values and interests of the trustees.

But if epistemic trust only refers to epistemic considerations (i.e. whether a trustee will produce true or epistemically well-founded outputs), why would alignment of values and interests be a relevant consideration? Surely interests and values do not pertain to epistemic trust but to some other kind of trust, like moral trust. Irzik and Kurtulmus's response stems from their analysis of inductive risk problems which, in short, point out that nonepistemic values play a role in determining the thresholds of evidence that must be met for the acceptance or rejection of hypotheses (Rudner, 1953). For example, determining that a water source contains "safe" toxin levels depends on the level of risk one is willing to tolerate (Douglas, 2000, 2009). In this way, a nonepistemic consideration is introduced to an epistemic process.

The challenge, however, is that nonepistemic values and interests often conflict. What one person sees as an acceptable level or risk may not align with another's. So, the idea behind enhanced epistemic trust is that if non-epistemic considerations (i.e., how to balance problems of inductive risk) play a role in epistemic processes, then those non-epistemic considerations should also factor into epistemic trust. For example, a parent might believe that a physician is well versed in all the risks and benefits of childhood vaccines, and in this respect, he holds basic epistemic trust in the physician to report true or accurate claims about vaccine efficacy and safety. However, the parent might also worry that the physician's idea of an acceptable level of risk associated with vaccination is higher than the parent's idea of what an acceptable level of risk is to subject his child to. In this way, the parent does not hold enhanced epistemic trust in the physician; the parent might be confident in the cogency of the physician's factual claims, yet still be reluctant to act on the physician's advice. This differentiation between basic and enhanced trust illustrates how an expert can be highly skilled and fully intend to use her expertise to honestly advise those seeking her aid (i.e. she will not lie or intentionally mislead), yet not attract as much epistemic trust as she would if she could demonstrate that her values and interests align with those of the trustor.

Accordingly, an ideally epistemically well-functioning system will have mechanisms in place to enforce epistemic standards that are not only truth conducive, but that also call for sensitivity to audience values to navigate problems of inductive risk where truth is underdetermined by available evidence. In the following subsections I propose rearticulations of John's sociological and epistemological premises to incorporate Irzik and Kurtulmus's notion of enhanced epistemic trust.[3]

1.1 The sociological premise

In analogue to John's sociological premise for epistemically well-functioning communities, my *sociological premise* for systemic trust rooted in epistemically well-functioning communities is as follows:

Sociological premise: The social-epistemic systems in which an expert is embedded are such that the best explanation for some expert's trustworthiness is

[3] John (2021) also takes inspiration from Irzik and Kurtulmus's (2018, 2019) notion of enhanced epistemic trust to make a similar modification to his original (John, 2018) two-step inference model. The main difference in our accounts lies in our motivation. My goal is to develop an account of "epistemically well-functioning systems" to articulate the conditions under which enhanced epistemic trust in experts is well-placed. John's (2021: 6) goal is to develop an analogous account of "sociologically well-ordered" epistemic communities to demarcate between scientific and non-scientific pursuits. The modification I propose to John's (2018) two-step inference was developed independently of John's own (2021) modification.

> that systemic factors at play effectively influence expert behavior such that it meets the epistemic and ethical standards of the field.

As discussed in Chapter 4, systemic factors are those features of the social epistemic systems in which experts are embedded that influence the individual's behavior. The systemic factors that underpin systemic trust in an expert roughly fall into two categories. First, *Extrinsic factors* are measures externally imposed on experts and expert communities to induce experts to behave in a certain way and maintain certain standards regardless of their desire to do so. Extrinsic factors – including policy, certification requirements, monetary incentives, education standards, social pressures, and audits – are explicit requirements and restrictions for behavior that outline consequences for noncompliance and/or provide external incentives for actors to behave (or not behave) in a certain way. For example, the consequence for a physician not keeping up with her continuing education requirements is the suspension or revocation of her medical license. Assuming the physician attaches some value to her license – e.g. social value if her professional identification grants her a high social standing in her community, or financial value in so far as practicing medicine brings in a hefty paycheck – then the threat of license revocation provides external social, and/or financial incentive to keep up with the latest medical research.

Second, *intrinsic factors* such as professional norms, cultural values and internal motivations cultivate within experts a desire to behave a certain way and maintain certain standards. Norms, for instance, are principles and values that exist within a social community that set expectations for, guide, or control acceptable behavior among community members. Sociologist James Coleman (1994) explains that community norms are most effective at influencing individual behavior when individuals internalize the norms such that an individual feels internally generated rewards and punishments. A social norm against littering, for instance, provides internal motivation for behavior if a person feels shame for dropping a candy wrapper in the park or pride or self-righteousness for picking up after others.

Overall, if extrinsic rules are thought of as the letter of the law, then intrinsic factors like norms and value can be thought of as the spirit. Not every possible action can be dictated by explicit restrictions and requirements, and where extrinsic regulation underdetermine behavior, cultural values and norms of conduct provide overarching and socially reinforced guidance and motivation.

I propose that the sociological premise for systemic trust is satisfied if, as discussed in Chapter 3, the networks of intrinsic and extrinsic systemic factors in which an expert is embedded do succeed in having a positive regulating influence on individual expert behavior such that it meets the epistemic and ethical standards of the field. In other words, satisfaction of the sociological premise requires fulfilling a success standard.

The question that follows, then, is about what kinds of intrinsic and extrinsic systemic factors are likely to yield such success. Specific guidance will depend on the scenario under consideration. The

conditions that will effectively regulate and ensure quality expert behavior will vary depending on the system, the people involved, and their interests and goals. Nonetheless, I can still comment on the more general point that the overall effect systemic factors have on individual conduct is resultant not only from the individual intrinsic and extrinsic factors at play, but how those factors interact and play off one another to constrain behavior and to harness actor motivations. I expand on this topic in Sections 3 and 4.

1.2 The enhanced epistemological premise

Even if the sociological premise is satisfied by the development and implementation of effective intrinsic and extrinsic measures within a system, it does not necessarily mean that systemic trust in an expert embedded in that system is well-placed. Epistemic safety requires that systemic factors enforce the right kind of standards. Toward this point, John put forward his epistemological premise. I expand on John's epistemological premise to incorporate both the epistemological and evaluative elements key to Irzik and Kurtulmus' notion of enhanced epistemic trust.

Enhanced epistemological premise: If systemic factors effectively influence an expert's behavior such that she meets the field's epistemic and ethical standards, then I should defer to the expert.

The enhanced epistemological premise adds to the sociological premise that systemic trust in an expert is well-placed if systemic factors effectively influence an expert's behavior so that it meets the field's epistemic and ethical standards *and* if those epistemic and ethical standards are such that I should rely on the outputs.

Like John (2018), I propose that satisfying enhanced epistemological premise requires the satisfaction of a success standard. Effectively implemented systemic factors ought to yield experts who reliably succeed in achieving the goals of the service they provide. Structural engineers will succeed in designing structures that remain upright, economists will most often succeed in predicting how market trends will respond to different events, and physicians will succeed in providing correct diagnosis and recommending effective treatments. My addition to John – also independently proposed by John (2021). See footnote 3 – is that the idea of what constitutes success (or acceptable risk) can have both epistemic and evaluative components. As such, in order to ground enhanced epistemic trust in experts, effectively implemented systemic factors ought to yield experts who act primarily to benefit those they serve (and inversely to not cause harm). Where there are problems of inductive risk such experts will respond to consumer values and interests in deciding how to navigate the uncertainty.

–

I have so far argued that systemic trust is well placed if systemic factors effectively influence individual expert behavior via intrinsic or extrinsic measures or some combination thereof (the sociological premise), and if those systemic factors promote epistemic and ethical standards which, when satisfied, are conducive to expert trustworthiness (the enhanced epistemological premise). In the following two sections I evaluate the sociological and enhanced epistemological premises of systemic trust in AI systems.

Note, I will not be commenting on how AI systems themselves respond to systemic factors, but on how the people and infrastructure employing and producing the AI systems do. It would admittedly be a stretch to discuss, for instance, how AI systems are motivated by internalized values or if they resent proposed regulations. However, as discussed in Chapter 3, systemic trust in AI, like systemic trust in human experts, can be analyzed from the perspective of the human led systems of production and systems of employment in which the AI system is embedded. In the context of employment, systemic trust in AI is rooted in networks of systemic factors that influence how, when, and by whom the AI system is used, while in the context of production systemic trust is rooted in the systemic factors that influence the research, design, and development of AI. In Sections 2 and 3 I appraise current foundations of systemic trust in AI with respect to systems of AI employment and production in turn.

2. SYSTEMS OF AI EMPLOYMENT

In the context of employment, systemic trust in a human expert is rooted in the systemic factors influencing the institution to which she is employed. For instance, systemic trust in a physician is rooted in the well-functioning of employing medical institutions. While, as discussed in Chapter 3, a patient is likely to have little evidence to substantiate idiosyncratic trust in the physician, the patient might still hold systemic trust in the physician because she believes the medical institution in which the physician is employed has measures in place to ensure only competent medical professionals are employed and to maintain standards of responsible conduct and patient centered care among practitioners.

Systemic trust in AI can likewise be rooted in the systems of its employment. For instance, one might trust an AI medical diagnostic system not because they know anything about the AI system itself (idiosyncratic trust), but because they believe the medical institutions that use the AI system has measures in place to ensure only reliable AI tools are employed and to ensure those tools are used responsibly with the aim of promoting patient well-being.

Recall from Chapter 3, however, that just because a person might hold strong systemic trust in a human physician does not necessarily mean one ought to also hold strong systemic trust in an AI system embedded in the same system of employment. One can have reason to trust an employing institution with

regard to one activity (e.g. hiring good physicians) but not necessarily with regard to another (e.g. employing reliable AI tools). This could be, for instance, because strategies differ for vetting human experts for employment than for vetting non-human instruments and tools. Accordingly, an appraisal of AI trustworthiness rooted in systems of employment must attend to the intrinsic and extrinsic factors that specifically influence how AI systems are chosen, maintained, and used within a given context of employment. Again, to be clear, I am not saying that AI systems can have intrinsic motivations. I only mean to discuss how intrinsic factors influencing AI employers might impact the epistemic trustworthiness of AI.

The foundations of systemic trust in contexts of employment will be context specific, and the contexts in which AI systems are employed are numerous including medicine, law, finance, transportation, and so forth. This makes it very difficult to conduct a thorough appraisal of current foundations of systemic trust in AI as rooted in systems of AI employment. Each context of employment will be guided by different standards which may or may not satisfy the enhanced epistemological premise. Furthermore, each context of employment will be characterized by different norms and values (intrinsic factors) and explicit rules and regulations (extrinsic factors) that determine how effectively the use and maintenance of the AI tool is regulated (how well the sociological premise is satisfied). As such, a thorough appraisal of all systems of employment that contribute to systemic trust in AI would take more space than I can offer here.

However, I can talk more generally about higher-level systemic factors meant to be applied across a variety of AI employment contexts as opposed to systemic factors deployed within individual use contexts. These might include, for instance, overarching requirements for record keeping, AI system maintenance, or human user training that could be applied where AI systems are used in a variety of settings such as aiding in medical diagnosis, criminal parole hearings, and finance management. I will comment on the European Commission's (2021) suggestions for high-level regulation of this sort shortly.

More importantly, there is reason to believe that these higher-level systemic factors are more important for establishing systemic trust in contexts of AI employment than lower-level systemic factors. By lower-level systemic factors I mean measures implemented closer to the individuals or products being regulated; for example, a checklist developed by a particular hospital to guide medical staff in the responsible use of AI diagnostic and treatment planning tools. It may be the case that such lower-level systemic factors have more direct influence on actor behavior and product use. However, just as it is often the case that a trustor will not have reason to hold idiosyncratic trust in an individual expert or AI system (see Chapter 3), I posit that a person will often not have reason to hold systemic trust in an individual based on what she, the trustor, understands about the lower-level factors influencing how AI tools are used and maintained. A trustor will most often not be aware of the specific rules or requirements

implemented by a trustee's employing institution. It logically follows that she will not be aware of what epistemic or ethical standards those rules aim to enforce (toward the satisfaction of the enhanced epistemological premise), and whether those specific rules are effective (toward the satisfaction of the sociological premise). Instead, she assumes that lower-level institutions are, in turn, responsive to even higher-level rules and regulations which work to ensure the necessary conditions obtain for enhanced epistemic trustworthiness. For example, a patient might trust a physician because the physician is employed by a GP practice, but she trusts the GP practice to employ competent physicians because the practice is, in turn, embedded in the larger National Health System which regulates the quality of medical services nationwide. As such, extrinsic measures employed at a high level (i.e. national or international) are worth a close look when evaluating the foundations of systemic trust in AI.

The concern I hold, however, is that there is not as much to say about high-level restrictions on AI employers as there should be. Most literature on trust and AI is concerned with the context of AI production. Common questions concern what design features are characteristic of trustworthy AI (Brundage et al., 2020), what high-level principles should guide AI research and development practice (Jobin et al, 2019), and how AI research and development should be regulated to ensure AI products are trustworthy, safe, and reliable (European Commission, 2020; Brundage et al., 2020). I will discuss systems of AI production in Section 3. Less attention, however, is paid to the context of AI employment and to the role systems of employment play in fostering user trust by ensuring a product is used responsibly and maintained appropriately.

The European Commission's (EC) (2021) Legal Framework for AI both illustrates this point and stands as a notable exception. The Legal Framework, which may be enforced as early as mid-2022, outlines a set of "harmonized rules for the development, placement on the market, and use of AI systems in the Union." The aim is to deliver on the EC's previously stated goal of building an "ecosystem of trust" around AI in which consumer of AI-enabled services can be confident that the technology is reliable, that the technology is used in a way that is safe and compliant with the law such that the consumer's fundamental rights are respected.[4] These rules are laid out in Title III, Chapter 3, articles 16-29 as obligations for various actors throughout an AI system's life cycle. Of those articles, 16-28 pertain to actors involved in the context of AI production – the development and distribution of AI-enabled technologies before they are placed on the market. These actors include AI providers – a person or entity "that develops an AI system or that has an AI system developed with a view to placing it on the market..." – and authorized representatives thereof, AI manufacturers, importers, and distributors.[5] Only article 29

[4] The goal of developing a policy framework to help in establishing an "ecosystem of trust" in AI is announced in the European Commission's 2020 white paper on trustworthy AI (European Commission, 2020).
[5] Definitions for each actor are provided in Title I, Article 3 of the Legal Framework.

stipulates obligations for AI employers.[6] Article 29 requires that AI employers follow instructions provided by AI providers, cease AI use when malfunction is suspected, report any problems to the provider, and keep the performance logs generated by AI systems. (Title III, article 12 requires AI providers to design high-risk AI systems with automatic record-keeping capabilities).

Though not yet enforced, the EC's Legal Framework is a large step toward building foundations for systemic trust in AI via systems of production and employment alike. The framework takes a systemic approach to establishing an ecosystem of trust around AI by addressing all aspects of an AI system's life cycle from is initial design and development through to its deployments and use, and it considers both epistemic factors (e.g. system reliability and proper use) and ethical factors (e.g. system safety and preservation of human rights) that would be necessary for the satisfaction of the enhanced epistemological premise. Furthermore, toward the satisfaction of the sociological premise, Title III, Chapter 4 and Title VI, Chapters 1 and 2 detail a multilevel system of oversight comprised of national competent authorities, notifying authorities, and an overarching European Artificial Intelligence Board to provide a mechanism for accountability, reporting, and enforcement.[7]

That said, one shortcoming of the EC's Legal Framework, which is neither the fault of the European Commission nor a problem with the content of the Legal Framework, is that the Legal Framework will apply only in the European Union leaving out countries such as the United States and China which compete to be world leaders in AI production and employment. Though these countries must meet EU standards to deploy their AI technologies inside the EU, outside the EU there is little foundation (current or projected) for systemic trust grounded in systems of AI employment. In this way, there can be greater foundation for systemic trust in any given AI system in one country than in another.

A second shortcoming, which does pertain to the Legal Framework's content, is the limited attention paid to the context of AI employment after an AI system is placed on the market. That greater emphasis is placed on regulating context of production is understandable. Production is chronologically prior to employment, and if quality AI products are not produced in the first place, working to ensure

[6] What I refer to as AI 'employers' the EC's Legal Framework refers to as 'users' where a user is understood as any person or entity "using an AI system under its authority, except where the AI system is used for personal-non-professional activity" (Title I, art. 3, para. 4). Users include AI providers who develop AI systems for their own use (e.g., a financial institution that develops its own financial market prediction tool to aid in investment decisions) and AI users who acquire AI products from external providers. (e.g., a hospital that purchases an AI medical diagnostic tool from DeepMind Health). I refer to such users as 'employers', as opposed to users, to avoid confusions with people who use AI systems for personal non-professional activities, who I would refer to as 'end users'. It is the end user's ability to ground her trust in an AI system that I am concerned with in this dissertation. If an end user were to acquire an AI system directly from a provider or to develop it herself, there would no context of employment for the end user to ground systemic trust in, only a context of production. Where the end user is using an AI tool provided by an employer, say the NHS who in turn acquired the tool from a provider, say DeepMind Health, the contexts of both production and employment come into play, DeepMind being a key actor in the context of production and the NHS framing the context of employment.

[7] Definitions are detailed in Title I, Article 3 of the Legal Framework.

those products are employed properly might seem like putting the cart before the horse. However, even the best made technologies can be used irresponsibly, unfairly, or put to nefarious ends. As such, attending to the context of AI employment is important too.

The EC's Legal Framework does cover employer responsibilities; however, these responsibilities are underemphasized. The EC's proposed measures stipulate that employing entities follow AI provider instructions, keep records, and report any problems to providers. But note how employer responsibilities amount to punting responsibility back to AI providers. Underpinning systemic trust in systems of employment is not just about ensuring AI employers follow user manuals and report any issues to providers. A well-functioning system of employment must also ensure human actors interact with the technology effectively. The introduction of a new technology into a social setting – like the employment of an AI enabled tool in a team – changes the dynamics of that setting. For example, Aimee van Wynsberghe and Shuhong Li (2019) describe how the introduction of "bots" (technologies ranging from embodied robots and AI to avatars and chatbots) into medical settings restructures healthcare systems. Namely, resources are allocated differently to account for purchasing and maintaining "bots", and the roles of human healthcare staff change as AI-enabled "bots" are delegated tasks traditionally fulfilled by human experts. Such changes in turn influence how medical decisions are made and how medical teams must interact to deliver quality care. If the acquisition of an AI component changes the social dynamics within a system of employment as van Wynsberghe and Li describe, then a well-functioning system of AI employment must also have measures in place to ensure this change in dynamic is managed well. To put it another way, if the employment of an AI tool (or any tool for that matter) causes changes to a system, then systemic trust is only maintained if the members of the system are given training to handle the change. In this respect the EC's Legal Framework for AI Regulation misses an opportunity to establish stronger foundations for systemic trust in AI by outlining education and training requirements for employers of high-risk AI technologies.

In summary, it is difficult to mount a detailed analysis of systemic trust in AI grounded in systems of AI employment due to the number and diversity of applications to which AI systems can be employed. However, higher-level systemic factors like state policy can be more readily analyzed. The challenge, however, is that most discussion about AI regulation pertains to contexts of AI production. The EC is a notable exception that includes legal stipulations for AI use in its plan to build an "ecosystem of trust" around AI systems deployed in the EU. Though I comment on how the EC's Legal Framework could do more to underspin trust in systems of AI employment specifically, I overall find it a promising start and a large step in the right direction, especially if the rest of the world follows suite.

I will now discuss foundations of systemic trust in systems of AI production. Here there is more fodder for discussion which will allow for more in-depth analyses of the enhanced epistemological and sociological premises independently.

3. SYSTEMS OF AI PRODUCTION

As introduced in Chapter 3, the context of production pertains to the initial development and training of a trustee. In the case of human experts, aspects of systems of production might include educational institutions and professional training programs, whereas in the case of AI, systems of production are comprised of systemic factors that influence the pre-market research, design, and development of an AI system. In this section, I appraise whether and how current systems of AI production might allow for well-placed systemic trust in AI.

3.1 Enhanced epistemological premise

In evaluating the epistemic well-functioning of systems of AI production I begin with the enhanced epistemological premise: are there epistemic and ethical standards in place in systems of AI production that would, assuming they are effectively implemented, render user epistemic trust in AI well-placed? The answer is both yes and no.

With respect to the "yes", there is no shortage of epistemic and ethical principles that have been proposed to guide AI research and development and which might work to satisfy the enhanced epistemological premise. Recent reports estimate that there are at least 70 publicly available sets of high-level principles, guidelines, values, or tenets for ethical and trustworthy AI design that have been proposed by governments, AI companies, and independent AI ethics initiatives (Morley et al., 2010; Floridi, et al. 2018; Jobin et al., 2019). Encouragingly, there seems to be some consensus among these guiding principles. Systemic reviews of ethical AI frameworks indicate that calls for accountability, fairness, transparency, explainability, safety, and user beneficence, well-being and autonomy are standard (Jobin et al., 2019; Hagendorff, 2020). Accordingly, we might assume that these high-level AI principles go at least some way towards satisfying the enhanced epistemological premise for enhanced systemic trust in AI. For instance, principles emphasizing AI safety generally speak to AI system performance (thereby speaking to the epistemic concerns), while principles centered on fairness, justice, and user autonomy and beneficence are generally meant to respond to user interests and to ensure AI systems are deployed with an eye to human well-being (thereby speaking to concerns about alignment with user interests and values).

However, (here comes the "no") while the intentions behind high-level AI principles suggest progress toward satisfying the enhanced epistemological premise, there is significant concern that high-level principles are too broad and not well enough defined to guide ground-level AI research and development activities (Mittelstadt, 2019; Peters et al., 2020). Take, for example, the principles of transparency and explainability. As discussed at great length in Chapter 1, the terms are often used interchangeably and there is significant confusion about if and how they ought to be distinguished from one another (See also Lipton, 2017; Tomsett et al, 2018; Krishnan, 2019). Even where it is clarified how explainability and transparency are distinct, other questions emerge: What purposes are AI explanations meant to serve? What kind of information should they contain? I argue in Chapters 1 and 2 that at least for the purpose of helping users to justify their belief in AI outputs, explanations should not be expected to achieve much. To complicate the issue further, the principles of explainability and transparency also do not always refer to desirable characteristics of AI products. These principles may also refer to the activities of individual AI systems, to AI companies, or even countries with respect to their AI development goals and strategies (Nyrup & Robinson, 2022).

Such lack of clarity around high-level AI principles underpin widespread concerns about "ethics washing" and unwarranted virtue signaling (Nemitz, 2018). As Brent Mittelstadt (2019) writes, "signing up to self-regulatory codes lacking clearly defined and enforceable obligations costs developers nothing but can have immediate benefits in terms of [perceived] trustworthiness and reputation." AI ethics initiatives based in general principles but not backed by executable plans serve as little more than smoke and mirrors to convey an image of trustworthiness (Maynard, 2019) and to convince policymakers there is no need to pursue enforceable regulation (Calo, 2017; Mittelstadt, 2019).

Clearly more work needs to be done to clarify high-level guiding principles, though there is reason to be optimistic on this front. A great amount of ongoing research is dedicated to better understanding what these principles mean and how to ensure their interpretation is inclusive of diverse human experience and concerns (*inter alia* see Hampton, 2021; Benjamin, 2019; Collett & Dillion, 2019). That said, clarifying high-level principles is only the first hurdle to building epistemically well-functioning systems of AI production; those principles must then be translated into practice (Peters et al., 2020; Morley et al., 2019; Zeng et al., 2019). This second hurdle falls under the sociological premise.

3.2 Sociological premise

Building foundations for systemic trust in AI requires that high-level principles are eventually embodied in ground-level AI research and development activities. Facilitating this outcome, however, is not a straightforward process. As Mittelstadt (2019) describes it, the "translation [of principles to practice]

involves the specification of high-level principles into mid-level norms and low-level requirements" and this translation takes place through complex processes of interactions and back-and-forth between intrinsic factors and externally enforced extrinsic restrictions. In what follows I take a closer look at hurdles to satisfying the sociological premise in terms of the intrinsic and extrinsic measures that serve to either support or thwart the translation of principles into practice. I begin with intrinsic factors.

As introduced in Section 1, intrinsic factors are factors such as professional norms and cultural values that emerge within communities and cultivate motivation within individual actors to behave according to certain standards. Accordingly, if a community of practitioners has internalized principles that satisfy the enhanced epistemological premise into community norms and values that succeed in guiding individual behavior, then there would be some foundation for systemic trust in that community and its outputs.

The medical profession is often taken as a gold standard in this respect, and various philosophers have engaged with the idea that trust in AI research and development might be bolstered by introducing principles analogous to Beauchamp and Childress's biomedical principles (Cave et al., 2021; Floridi et al., 2018; Mittelstadt 2019; Whittlestone et al., 2019). It is not clear, however, that the same success with principle-based regulation might be achieved within the context of systems of AI production.[8] Brent Mittelstadt (2019) worries that the ways in which the medical profession is (or may be) well-situated to facilitate the internalization of guiding principles within a community of practitioners, the AI research and development community differs.

As discussed in the previous section, Mittelstadt notes the challenge of defining principles with sufficient clarity and then translating those principles to practice. Toward this end, the medical profession has benefited from a long history in which medical principles have evolved and been refined, and in which infrastructures have been constructed to facilitate the integration of stated principles into professional norms and practice. Modern medical norms did not emerge overnight. For example, contrary to popular belief, the medical mantra "do no harm" was not embodied in Western medicine as soon a Hippocrates drafted his Oath. Neither did Beauchamp and Childress's biomedical principles centered on patient understanding and autonomy instantly supplant the paternalistic tradition of, "doctor knows best" internalized by our grandparents. But we can point to key extrinsic measures that ferried the process along (Brothers et. al., 2019). For example, the Nuremberg trials, leading to the publication of the Nuremberg

[8] There is also debate as to whether high-level principles can and should be used to help guide even medical practice. For example, following on Beauchamp and Childress's (1979) popular articulation of the four biomedical ethics principles, Clouser & Gert (1990) lead a cautionary opposition arguing that "at best, 'principles' operate primarily as checklists naming issues worth remembering when considering a biomedical moral issue. At worst 'principles' obscure and confuse moral reasoning by their failure to be guidelines and by their eclectic and unsystematic use of moral theory" (220). Also see Davis (1995).

Code (1947), heavily stigmatized medical experimentation and/or treatment without patient consent, as did the Tuskegee syphilis trials and other trials described by Henry Beecher (1966). Later, the Belmont Report (National Commission for the Protection of Human Subject, 1979) and Childress and Beauchamp's (1979) *Principles of Biomedical Ethics* outlined biomedical principles centered around themes of patient autonomy and beneficence now taught throughout medical training and reinforced through explicit legal requirements. AI development, on the other hand, is a comparatively recent endeavor. As Brent Mittelstadt writes:

> AI development does not have a comparable [to medicine] history, homogenous professional culture and identity, or similarly developed professional ethics frameworks. The profession has not gone through comparable transformative moments in which its ethical obligations are clearly recognized and translated into specific, practical moral duties and best practices (Mittelstadt 2019: 503).

However, if the comparatively short history of AI development as a profession is the crux of the problem, then a solution should follow with time. But the challenge, Mittelstadt argues, is more complicated than this.

Mittelstadt contends that the diverse backgrounds and professional educational experiences of AI researchers poses a significant barrier to the widespread internalization of high-level principles and norms of conduct. Professions can generally be understood as organized groups of practitioners characterized by their specialist knowledge, formal education, training requirements, regulated activity (professional practitioners are licensed to practice), codes of ethics, and ethos of public service. Practitioners within a specific profession are also traditionally bound by a common set of norms and values (Iacovino, 2002; Susskind & Susskind, 2016). The worry is that unlike, say, medical professionals, "AI developers come from varied disciplines and professional backgrounds, which have incongruous histories, cultures, incentive structures and moral obligations", and as such, we should not expect AI researchers and developers to develop the kind of common professional identity that is conducive to the uniform internalization of high-level principles (Mittelstadt, 2019: 503). In other words, we should not expect that AI researchers and developers will ever respond to the regulatory effects of high-level statements of principles in the same way communities of physicians might be expected to.

If intrinsic factors cannot be relied upon to guide the development of safe and beneficial AI, extrinsic measures – explicit rules and requirements – must be used to drive desired behavior. Accordingly, Mittelstadt suggests requiring licensure for people and companies developing high-risk AI technologies. Only parties who fulfill specific requirements for record keeping, conducting risk analyses, meeting security or privacy standards etc. will be able to compete in that market space. Another option is to provide direct monetary incentives, for example, to employees who complete AI ethics training courses and engage in continuous education programs. When such carrot-and-stick appeals to self-interest cannot

be used to incentivize desired behavior, dictated rules and requirements come into play. The European Commission's Legal Framework for AI regulation (detailed in the previous section) is a groundbreaking initiative in this respect. While not yet legally enforced, the framework is the first of its kind, laying out explicit constraints and responsibilities for developers, distributors, and importers of high-risk AI technologies, and it is likely to set a precedent and be a model for AI regulation globally.[9]

The merit of extrinsic regulation for satisfying the sociological premise and enforcing standards that satisfy the enhanced epistemological premise is undeniable, especially in the short term as AI development is couched against the Silicon Valley tech scene which prioritizes optimization, scale, and efficiency and is dominated by market motivations (Thompson, 2019: 21-23). Explicit rules are needed to ensure ethical guidelines and responsibilities are taken seriously. I caution, however, that it is unwise to put too much faith in the regulatory power of extrinsic measures alone. There are limitations to a heavily constraint-based approach to AI governance like that Mittelstadt recommends.

First, it is difficult to articulate precise rules for AI research and development which is conducted for a wide array of applications and the field evolves quickly. For example, the EC's Legal Framework for AI requires risk management systems to be established for all high-risk AI applications (Title III, Art. 9). General features of such risk management systems are provided, though specific details on how risk assessments should be conducted for AI applications being deployed to different industries are lacking (Title III, Art. 9, Para. 2-4). The exception is with regard to AI deployed to credit institutions, in which case the EC framework defers to the specific risk management procedures articulated in those institutions' independent directive on AI risk analysis (Title III, Art. 9, Para. 9). A key insight here is that articulating any sector specific and practicable requirements for AI research and development will require the involvement of the intended AI employers. In this respect it is clear that systemic trust in AI will be rooted both in systems of AI production and systems of AI employment (Section 2) and in the interactions therebetween.

Second, even where explicit rules and procedures can be articulated, extrinsic regulatory measures are more effective as tools for ensuring that the sociological premise is satisfied if they are supported by the intrinsic norms and values already internalized by a community. Seth Baum (2017) exemplifies this point by comparing public responses to a constitutional ban on flag burning and legal requirements for dog waste disposal in the United States. On several occasions the United States government has considered a constitutional amendment to ban flag burning. On each occasion the amendment has been rejected, one reason being that "adopting a constitutional amendment may be the best possible way to promote the incidence of flag burning" (Sunstein, 1996: 2023). According to current

[9] If the European Commission's 2018 implementation of the General Data Protection Regulation (GDPR) is anything to go by, the European Commission's Legal Framework of AI Regulation will be another example of the EU and UK's pioneering and globally influential work in tech regulations.

cultural norms, most people view flag burning as distasteful, disrespectful, and unpatriotic. However, the criminalization of flag burning could also be seen as an infringement of citizens' rights to freedom of speech and expression, rights in which patriotic Americans take great pride. For some, flag burning could become an act of patriotism in of itself and instances would increase. In contrast, recent laws requiring dog owners dispose properly of pet waste in New York City (Krantz et al., 2008) and Berkley (Cooter, 2000) have had great success in cleaning up city pavement without increased surveillance or fines issued by law enforcement. Rather, community members became more vocal about calling out litterers, and neighborhoods more effectively policed themselves. Baum's point is that extrinsic measures are more effective at regulating behavior when they are policies that are in accordance with intrinsic norms and values already internalized by a community. In this respect, extrinsic AI regulation should strive to be less like rules against flag burning and more like dog waste disposal laws. They should be measures AI practitioners would want to follow.

In terms of understanding the concept of extrinsic measures people "want to follow" and how such measures are best implemented, sociologist Julian Le Grand (2003) describes the efficacy of extrinsic rules and enforcement strategies as responding to how the subjects of the extrinsic measures are intrinsically motivated. Le Grand differentiates between knaves – individuals primarily driven by self-interest – and knights – individuals primarily driven by altruistic motivations and internalized social norms and values. The distinction is important if one wishes to predict how people are likely to respond to extrinsic rules and enforcement mechanisms.

For example, Le Grand warns against regulatory strategies that assume individuals are knights. Traditional professions, for instance, are defined in part by the supposed commitment of practitioners to a common set of values and behavioral norms. If the practitioners truly have internalized prosocial norms of public service and responsibility to the consumer (norms which would help underpin the enhanced epistemological premise by promoting sensitivity to consumer values and interests), then it may be acceptable to allow professionals a large degree of individual freedom. However, if some of the practitioners are influenced by knavish motivations, self-regulation could be dangerous. If consumers of professional services are not well equipped to identify threats to their own interests and to defend themselves, as will often be the case with lay consumers, self-interested knaves may take advantage of naïve consumers to serve their own interests at the consumer's expense.

Accordingly, it may seem that it is never advisable to allow experts (or AI producers) to self-regulate. Rarely, if ever, will a person be entirely influenced by knightly motivations, and therefore it would be wise to always err on the side of implementing rules and strict enforcement structures aimed at the regulation of knaves. As Le Grand writes, "In a situation of ignorance concerning the motivational structure of individuals, it would be safest to adopt public policies based on the assumption that everyone

is a knave. For a knavish strategy will do little harm if people are actually knights; and it could pay off dramatically if they are knaves" (53).

However, Le Grand continues, while it is dangerous to regulate knaves like knights, we should also be careful not to regulate knights like knaves. The knightly, 'other-oriented' mindset held by many professionals is a strong motivator for good behavior, and tight extrinsic regulation can be taken as an insult, an undervaluing of one's altruistic actions, or an inconvenient and unnecessary restriction on one's (deserved) freedom. Studies show that appeals to a professional's prosocial values (where they are held) is often more effective at motivating individual performance than carrot and stick incentive structures or explicit requirements (Anik, 2013). Le Grand explains that when knightly individuals are, for example, regularly audited they may rightly feel that they are no longer trusted, become resentful of the unnecessary extra work imposed on them, and, in turn, become less committed to the service they provide and more inclined to purse their own interests. "They become demoralized and demotivated or else motivated to behave in a more self-protective manner" (Le Grand, 2003: 57).

Philip Pettit (2002) makes a similar point in the context of biomedical research. Pettit warns that overly assertive and conservative restrictions passed by biomedical ethics committees may engender a culture of resistance among researchers who feel that potentially beneficial research is being unnecessarily prevented. He writes, "if researchers do come to lose a commitment to ethical guidelines, if they do come to be 'demoralized' this way... it may not only lead to a restriction of the research we currently tolerate. It may also lead to a restriction in the commitment of researchers to the ethic that currently prevails." We can similarly conclude that numerous time-consuming checklist requirements, like filing reports and filling out evaluations and impact statements, are likely to foster resentment even among knightly AI researchers and developers if the exercises are seen as unnecessary or superfluous. Involving workers in the development and enforcement of the extrinsic measure to which they are subject may help to prevent such resentment (Orentlicher, 2002; Iacovino, 2002; Schienke, 2009; Baum, 2017). I discuss this strategy further at the end of this section. But for the time being, the key point is that extrinsic measures will be more effective (more willingly taken up and adhered to) when practitioners are invested in their own regulatory structures and share in the values and goals which extrinsic measures are meant to promote.

This of course leads us back to the problem Mittelstadt originally pointed out, that a lack of professional cultural unity among AI researchers and developers poses a problem to the satisfaction of the sociological premise. But now, not only is disparate background and cultural disunity among AI practitioners a barrier to the internalization of guiding principles into intrinsic regulatory factors like social norms and internal motivations, that lack of internalization is also a barrier to effective extrinsic regulation. We seem to be faced with a vicious cycle: extrinsic measures are key to the effective

regulation of AI research and development because of a lack of professional unity among AI practitioners, but the lack of professional unity also threatens to lessen the efficacy of any extrinsic measures employed because there are no common community values with which extrinsic measures can align.

I think, however, that Mittelstadt overstates the challenge of cultural disunity among AI researchers in a couple ways. First, cultural disunity is not as bad as he makes it out to be. Rather, there is substantial evidence of a budding cultural identity among AI researchers centered on the high-level principles discussed previously which aim to satisfy the enhanced epistemological premise. Hayden Belfield (2020) argues, for instance, that the growth of cultural unity within the AI research and development community has been key to the success of a slew of recent AI activism activities which have aimed to promote the development of safe and beneficial AI, to encourage safe and responsible AI use, and to promote awareness of the greater societal implications of various AI applications. The emergence of a common professional identity centred on these "knightly" values is further illustrated in the move by major AI research conferences AAAI, NeurIPS, ICML, and IJCAI to include sessions on AI ethics and social impact. Furthermore, several major AI companies are strongly committed to operationalizing core AI principles. For example, OpenAI is an AI company established with the stated mission of ensuring that artificial intelligence benefits all of humanity and of aiding others to achieve this outcome. Microsoft has similarly deployed an arsenal of social strategies to nurture a company culture of product quality and responsibility which I will discuss in more detail shortly. Also notable is the Partnership on AI to Benefit People and Society (established September 2016) which now brings together over 90 AI companies and organizations to collaborate on establishing best-practice recommendations for the AI community. Clearly, despite the diverse backgrounds and educational/training experience of AI talent, a common professional culture centered on key AI principles is beginning to take shape.

My second response to Mittelstadt is that while I agree that a common professional and cultural identity does facilitate the internalization of high-level principles and the effective implementation of extrinsic regulatory measures, it should also be recognized that high-level principles and specific extrinsic measures can also be used to nudge cultural identities to evolve in a desired direction. For example, I have argued elsewhere (Seger, 2022) that many complaints against the use of high-level principles as tools of AI governance – e.g. complaints against ethics washing and the difficulty of translating principles into practical instruction – target only what I call the "start-point function" of high-level principles. In terms of their start-point function, principles are viewed as providing a departure point for articulating more precise rules and requirements to guide AI research and development. I propose, however, that a less discussed, though no less important role of AI ethics principles is in underpinning cultural norms and values. I call this the "cultural influence" function of principles. When viewed as a guide to cultural change and value reinforcement, it does not so much matter how exactly high-level AI principles are

delineated or defined. Irrespective of their translation into precise rules and regulations, high-level principles may serve as attention grabbing concepts that can be used to reenforce knightly values in the forefront of practitioner minds, and thereby to guide the evolution of intrinsic factors, cultural norms, and values. For instance, a recent white paper released by the World Economic Forum (2020) identifies a key function of high-level AI principles as providing a common vocabulary with which AI developers discuss design challenges and contemplate potential impacts and risks. That common vocabulary frames how practitioners construe the design challenges they face which, in turn, influences the solutions they entertain. Overall, many proffered AI principles such as fairness, accountability, explainability, inclusivity, and transparency challenge the status quo Silicon Valley tech culture focused on efficiency, optimization, and scale. The introduction of and subsequent debate over high-level AI principles are, as discussed in the previous paragraph, an instigating factor in the cultural shift toward the prioritization of safety, responsibility, and human beneficence (key to the enhanced epistemological premise) instead.

Cultural change is, of course, a complicated process. Encouraging AI developers to internalize new values and to alter their social norms to align with policy goals and consumer well-being is not merely a matter of posting a set of principles on a wall. Extrinsic rules and requirements also play an important role in reinforcing intrinsic values and 'nudging' norms to evolve.[10] Top-down extrinsic measures (rules and guidelines dictated by a higher authority) such as education and training requirements can be used to reinforce norms of service and commitment to user welfare from the start of an AI developer's professional development. Common experience in foundational stages of professional development helps set a groundwork on which a community of practitioners can form a shared identity based not only on their area of expertise, but also on their shared experiences and on the values communicated during that training. For example, upon one's employment by an AI development company, onboarding training, codes of conduct, and mission statements centered on high-level AI principles may be used to set the tone of the company's culture and indicate that company leadership takes seriously specific principles and values; employees are expected to keep the values in mind and conduct themselves accordingly. After onboarding, frequent reminders such as employee-evaluations, training days, and record keeping requirements structured according to high-level principles might also be used to keep AI principles center stage in AI research and development activities.

[10] The formation and enforcement of norms in a social setting is a complicated process that I cannot explain fully here. I direct interested readers to Coleman (1994: chapters 10 and 11) for a thorough analysis. In short, Coleman describes the emergence of norms as, "a prototypical micro-to-macro transition, because the process must arise from individual actions, yet a norm itself is a system-level property which affects the further actions of individuals" (244). In turn, individual actions reinforce the norm and so on and so forth. So, norms emerge, evolve, and persist through an iterative process of micro-social and macro-social reinforcement of social values. In this way policies and incentives can alter individual actor behavior and values which in turn feeds into the iterative 'micro-to-macro transition' by which norms take shape.

Microsoft's Responsible AI Initiative provides an illustrative example of how such top-down measures might be structured within an AI company.[11] Microsoft launched its top-down initiative starting by putting forth the company's own set of six principles for AI development – fairness, reliability and safety, privacy and security, inclusiveness, transparency, and accountability – with the aim of promoting cultural change and ethical awareness within the company.[12] The Aether Committee, an expert scientific and engineering advisory board, counsels on the implementation of the company's principles. The six principles are also accompanied by the Responsible AI Standard, a set of steps that AI research and development teams are required to follow in order to support the production of responsible AI systems according to the principles. A key requirement of the Responsible AI Standard is that any work on sensitive use cases of AI (cases in which AI applications might violate human rights or safety or are applied high-risk sectors like medicine or transportation) must be overseen by the Office of Responsible AI (ORA). The Standard was initially piloted among ten engineering teams, and feedback was collected to guide improvements for its next iterations. The Responsible AI Standard 2.0 provides more explicit implementation methodology, and all employees are required to undergo training that covers the Responsible AI Standard and the six principles.

It is important to remember, however, that the implementation of top-down measures alone can backfire and have undesired consequences for the sociological premise if not implemented carefully. As previously discussed, when extrinsic requirements are felt to introduce seemingly superfluous work, resentment is likely to fester. This resentment may, in turn, lead to rejection of the very principles and values that are being promoted. It is therefore advisable to involve those being regulated in the development and communication of extrinsic measures to avoid community wide demoralization and resistance (Orentlicher, 2002; Iacovino, 2002; Schienke, 2009). Expert involvement in formulating extrinsic measure and communicating norms of conduct to more junior practitioners is key to ensuring novices benefit from expert insights, to fostering a sense of moral responsibility within the community, and, overall, to promoting the uptake and efficacy of extrinsic regulations. As Seth Baum (2017) writes, using in-field experts as messengers "shows that ethical and social impacts is something that 'we' (i.e., people in the field) care about and is not just something that 'they' (i.e., people outside the field) want

[11] Much of what I describe about Microsoft in the coming paragraphs is drawn from the World Economic Forum's recent white paper *Responsible Use of Technology: The Microsoft Case Study* (WEF, 2021). It is first in an upcoming series of investigative papers examining how companies have begun to incorporate ethical thinking into technology development. This series will be an important space to watch for those interested in better understanding the real-world benefits and pitfalls of tech company self-governance. Where I draw from other sources, I make a note.

[12] Microsoft, *Responsible AI,* including videos, 2020.
https://www.microsoft.com/en-us/ai/responsible-ai?activetab=pivot1:primaryr6 (accessed April 7, 2021).
Microsoft. *Responsible AI Resources*.
https://www.microsoft.com/en-us/ai/responsible-ai-resources?activetab=pivot1%3aprimaryr4 (accessed April 7, 2021).

'us' to care about. This is reflective of a more general tendency for people to respond better to messages from other people in their 'in group'" (548). On this point, in the medical context Pettit (2002) recommends against having disconnected groups of "ethics enthusiasts" laying out guidelines to regulate the efforts of separate groups of physicians and biomedical researchers.

Ethics regulations and principles will be less effective if dictated from on high by external 'ethics enthusiasts'. Accordingly, to help satisfy the sociological premise for well-functioning systems of AI production, it would be wise to involve AI workers in any ethics boards or internal review committees to avoid ungrounded virtue-signaling and to encourage community investment in promoting desired standards of conduct. For example, Microsoft accompanies its top-down rules and requirements for responsible AI with infrastructures to facilitate bottom-up involvement in AI governance. Key among them is the Responsible AI Champs program. "Champs" are experts within Microsoft's AI design and development teams who also work to help fellow employees identify ethical challenges in their work, to point colleagues toward available responsible AI resources, and to otherwise foster a culture of responsibility within the company. In other words, Champs work as full-time advocates and evangelists for responsible AI.[13] Microsoft also has a Responsible AI Strategy in Engineering (RAISE) team that works in collaboration with Microsoft engineers to build tools and resources to help engineering teams implement the company's AI principles and internalize the values the principles embody.[14]

Overall, without bottom-up expert involvement, top-down efforts to introduce and reinforce high-level principles will struggle to find traction. On the other hand, top-down involvement from company leadership like the oversight provided by Microsoft's Office for Responsible AI and Aether committee is needed to provide the resources and support necessary for bottom-up initiatives like RAISE and Champs to succeed. Top-down support allows utilization of established organizational infrastructure and monetary resources and offers validation for the values and agendas being pushed. Without it, grassroots initiatives will struggle to impact the quality of AI products and are likely to lose momentum.

It is, of course, yet to be seen whether Microsoft's efforts do indeed result in AI products which are more epistemically trustworthy – that is, more likely to produce epistemically well-founded outputs that respond to user interests and values – than before Microsoft launched its Responsible AI Initiative in 2019. My point is rather that Microsoft's approach to dual top-down and bottom-up regulation of AI production is one which, in theory, we might expect to underpin systemic trust in AI products.

[13] See footnote 11.

[14] For example, the Ethics & Society team with RAISE has developed several responsible innovation tools that aim to "exercise the moral mind". These tools include harms modeling interfaces, community jury techniques (adapted from citizen juries) and the Judgement Call game – an interactive role-playing game in which AI developers take the perspective of different stakeholders to explore potential impacts and harms.

There are, however, an assortment of challenges to thinking of Microsoft's model as one that might be emulated more widely to underpin systemic trust in AI. First, AI research and development is not a project isolated to large companies endowed with the kinds of monetary and human resources enjoyed at Microsoft. If Microsoft's model does work, it is because it has been deployed within Microsoft, but small companies and startups that lack similar financial and human resources cannot be expected to undertake similar efforts.

Second, responsible AI initiatives instigated by individual companies or research organizations are prone to instability. It is reasonable to assume that the activities of private AI producing companies are primarily driven by market incentives. So, where knavish financial self-interests conflict with potentially costly (in terms of financial and human resources) initiatives to enforce responsible and socially beneficial AI development, some extra driving force is needed to push those initiatives through their fledgling stages. However, when that initial knightly push emanates from the willful persistence of an enthusiastic CEO or executive team (as in the case of Microsoft) it provides a precarious foundation for systemic trust. Consider, for example, the spontaneous dissolution of DeepMind Health's AI ethics initiative. In 2016, DeepMind Health Co-founder Mustafa Suleyman championed the establishment of an internal review panel to scrutinize the company's AI development and data-management practices (DeepMind Health, 2017). Despite the misgivings of DeepMind's board of directors, for the two years that the internal review panel existed, at Suleyman's insistence the committee was granted a large budget of £50,000 per year for commissioning external audits on the company, allowed unfettered access to company records and employees, and held to no confidentiality requirements in their communication with the press or in their publication of findings and recommendations. However, after the publication of the panel's second annual report (DeepMind Health, 2018), a junior journalist falsely reported the committee's recommendations to DeepMind Health as a vote of no confidence.[15] Though the journalist later retracted the statement and issued an apology, DeepMind's board of directors – already hesitant about the panel's freedoms – pulled the plug on the initiative.

We can expect Microsoft's program to be more robust than DeepMind's given their efforts to encourage widespread internalization by employees at all levels of the company via both bottom-up and top-down measures. Microsoft's program is not as liable to collapse given a change in leadership or a one-off media gaffe because Microsoft aims to build a community of researchers united by shared norms rather than indefinitely relying on the extrinsic dictates of an oversight body. However, Microsoft's efforts are isolated to Microsoft, and therefore, to the extent those efforts underpin systemic trust at all, they narrowly underpin systemic trust in Microsoft products alone. Furthermore, it will often be the case that a

[15] Skype interview with DeepMind Health Internal Review Panel chair, Julian Huppert (April 7, 2020).

lay user does not know by whom an AI system was produced and/or about the producer's efforts to ensure AI product quality. Higher level regulatory efforts are needed to ground systemic trust in AI more widely.

Note the parallel here to Chapter 3's discussion on idiosyncratic vs. systemic trust. I argued that most often the trust people hold in experts and AI systems is not based on their beliefs about the individual's (expert's or AI's) idiosyncratic characteristics but in their assumptions about the systems of employment and production (education/training) in which the individual is embedded. This is because people will often not have access to or not be able to use information about an individual's idiosyncratic characteristics to inform her decision to trust the expert/AI system. Here I make a similar observation that systemic trust in AI is likely to be grounded in higher-level systemic factors that influence the context of production than lower-level systemic factors. As described in Section 2, lower-level systemic factors are those intrinsic and extrinsic measures implemented closer to the individuals or products being regulated, like, for instance, the top-down and bottom-up measure implemented by Microsoft as part of its responsible AI program. Higher-level systemic factors, on the other hand, are further reaching. They include, for example, legal requirements and broader cultural norms that influence AI production on a wider national or international scale. If a user does not know by whom an AI system was produced and/or is unaware of lower-level systemic factors at play influencing product quality, then her systemic trust in an AI product may instead lie in her belief that there are higher-level systemic factors at play satisfying the sociological premise – that is, the belief that AI researchers and developers and the companies for which they work are part of higher-level networks of systemic factors keeping knavish corporate motivations in check and maintaining standards of AI production writ large. Overall, in an ideally well-functioning system of higher-level systemic factors, individual AI producing companies will be interchangeable with respect to the systemic trust one might hold in their products.

This means, for instance, that the internalization of norms of conduct and cultural values that satisfy the enhanced epistemological premise must be ubiquitous, not just a happy occurrence within certain organizations. In an ideal world, AI developers will not just be united in their identities as "Microsoft employees" or "Googlers" but more generally as "AI practitioners" just as physicians do not just think of themselves as Sloan-Kettering physicians or Mayo Clinic physicians but simply as medical doctors. It is this broader professional classification that must ultimately carry with it the mindset of professional service, fiduciary duty, and commitment to quality and consumer beneficence that will serve to underpin enhanced systemic trust in AI. Toward this end, initiatives that aim to establish a unified professional identity among AI workers are key. Such schemes could include standard education requirements at the university and graduate level and training requirements for professional qualification. For example, the European Commission's (2020) white paper on trustworthy AI recommends establishing "an ecosystem of trust," starting by setting standards for excellence in AI education and establishing a

centralized certification framework for AI developers operating in the EU. Continued reinforcement of professional values is also important. Toward this end, key journals and conferences in AI fields could, for example, require that impact statements accompany each submission such that work conducted without analyses of the potential benefits and harms of the specific AI developments are stigmatized (Ashurst et. al., 2020). AI researchers and developers could also be encouraged to pursue membership to professional associations like IEEE and AAAI that promote responsible AI development practices.

So, what does this all mean with respect to the foundations of systemic trust in the context of AI production? Are systems of AI production epistemically well-functioning such that they are likely to yield AI systems in which enhanced epistemic trust would be well-placed? To summarize, I have argued that we might assume the enhanced epistemological premise of well-functioning systems is at least on its way to being satisfied. There is general consensus surrounding well-researched AI principles that address both the epistemic and value-alignment concerns of the enhanced epistemological premise. However, as clearly exemplified in Chapters 1 and 2 with respect to AI explainability requirements, more work is needed toward definitional clarity to pave the way for the translation of high-level principles into practice.

The subsequent challenge is with the sociological premise – with implementing those principles via the manipulation of intrinsic and extrinsic factors such that they effectively influence AI producer conduct. There is significant concern, however, about the hurdles posed to effective regulation by the disparate backgrounds of AI practitioners and current cultural disunity within the AI research and development community – a lack of common values and goals among AI developers threatens to undermine the efficacy of extrinsic regulatory measure and any self-regulation initiatives. That said, and on a more positive note, there is ample evidence to indicate the emergence of a unified professional identity among AI practitioners which will continue to solidify with time. More so, there are steps that can be taken to facilitate the process. Namely, extrinsic measures and high-level principles can be used to nudge social norms and cultural values to evolve to support the satisfaction of the sociological premise. It is important, however, that the efforts to manipulate intrinsic and extrinsic factors are not just made at the level of individual companies and organizations as they are currently. Overarching higher-level efforts are needed to ground systemic trust in AI generally, that is, in the same way as there is general, baseline level of systemic trust in physicians irrespective of their specific places or training or employment. That said, as of now there are no explicit requirements of all AI producers beyond the General Data Protection Regulation's (GDPR's) stipulation that it must be possible to provide subjects of autonomous decision-making technologies – of which AI-enabled expert systems are a subset – intelligible explanation for how the system derived its outputs.[16] However, the European Commission's Legal Framework for AI

[16] I am, of course pessimistic that attempts to make AI output derivation processes transparent to lay users can help users make epistemically well-informed decisions about whether to believe AI outputs (see Chapters 1 and 2). That said, with respect to how AI explainability requirements might help underpin systemic trust, I am less pessimistic.

Regulation outlines a promising foundation for future regulation that spans various levels of AI production activities from initial research and design to product distribution and maintenance. The Regulation's stated goal is to build an "ecosystem of trust" around AI.

Overall, at the time I write this dissertation there seems to be little foundation for enhanced systemic trust in systems of AI production, but I propose that we can be optimistic about the future. Wheels are turning in the right direction.

4. CONCLUSION

This chapter concludes Part 2 of this dissertation on foundations of trust in AI and experts. Chapter 3 argued that both human novice-expert trust relationships and user trust in AI are best understood in terms of systemic trust as opposed to idiosyncratic trust, and Chapter 4 presented a brief interlude on my use of trust terminology with regard to AI systems and other artifacts. This chapter returned to the topic of systemic trust, first describing the requirements of an epistemically well-functioning system – one that satisfies the sociological premise and the enhanced epistemological premise – and second, investigating the epistemological well-functioning of systems of AI employment and production. While I argued that neither systems of AI production nor employment provide a strong basis for enhanced systemic trust in AI at this time, I observed that positive steps have been taken in both cases to start building more epistemically well-functioning systems around AI.

Before I wrap up, I need to return to the discussion of internal and external justification. This chapter has been about describing conditions under which enhanced epistemic trust in an AI system (or human experts) would be *externally* justified. In terms of epistemic security, this means ensuring people place trust in epistemically trustworthy sources by making the world epistemically safe such that epistemically untrustworthy sources are hard to come by. We must not lose sight, however, of the original question this dissertation aimed to address – how can people acquire a degree of *internal* epistemic justification for their belief in AI and expert outputs? That is, how can people determine for themselves whether a claim or output ought to be believed or whether a claim/output source is epistemically trustworthy? Perfectly epistemically safe systems are an unachievable ideal. As such, epistemic security must also be understood in terms of the sensitivity of individual members of society to truth and epistemic trustworthiness.

Chapters 1 and 2 show that AI explainability is unlikely to help with satisfying the epistemological premise; the explanation itself will not help the user determine whether the system employs epistemically sound methodologies. However, that a company satisfies requirements for building AI systems that produce explanations speaks to the satisfaction of the sociological premise; irrespective of how justificatorally valuable an explanation can be to a user, that the AI producer demonstrates rule following and the satisfaction of policy requirements might provide some evidence that the AI production processes is, more generally, responsive to extrinsic regulation and that it is functioning in the way it ought to be.

Toward this end, in Part 1 I explored the possibility of people epistemically justifying belief by looking at the content of claims or explanations thereof for coherence with their own background knowledge and beliefs. I came to the conclusion, however, that the more epistemically imbalanced a relationship the less justificatorally valuable an explanation can be. This means that in expert-novice and AI-lay user relationships, explanations should not be expected to provide users with substantial internal epistemic justification for believing expert claims or AI outputs. So, in Part II, I turned to epistemic trust. In Chapter 3 I explained that where that internal epistemic justification for belief in an output cannot be acquired by directly appraising the output or an explanation thereof, that internal epistemic justification for belief might instead be rooted in one's belief that the output was produced by an epistemically trustworthy source. Because novices and AI users are unlikely to be familiar with the idiosyncratic characteristics of the individual experts and AI systems they consult, I argued that epistemic trust in experts and AI systems is mainly systemic trust. Systemic trust is therefore trust based on one's beliefs about the presence of systemic factors that serve to regulate and maintain quality expert/AI performance. In this chapter, I then presented the sociological and enhanced epistemological premises which, when both satisfied, describe an ideally epistemically well-functioning network of systemic factors. The problem, however, is that in discussing the foundations and structure of epistemically well-functioning systems I have only been discussing what conditions must obtain for systemic trust in experts or AI systems to be externally justified – that is, objectively well-placed relative to who or what is indeed likely to be epistemically trustworthy source of information.

Internal justification for systemic trust requires, however, that users might evaluate for themselves whether networks of systemic factors are epistemically well-functioning. A trustor would have to be aware of the various systemic factors that are in place, and her decision to trust would have to be based in a well-founded belief that those systemic factors do positively influence trustee performance. It is not enough that a network of systemic factors is epistemically well-functioning. The presence and efficacy of those systemic factors also have to be indicated to relevant audiences in some way. For example, with respect to systemic trust in AI, new tech legislation should be clearly explained and publicized. Furthermore, certification programs could be set up for AI producers whose employees have undergone some standardized training and who implement safety and quality check practices. Similarly, AI employers could be certified as having undergone training in the safe and responsible use of the AI application they employ.

It is important to note that such efforts to flag systemic trustworthiness can serve a dual function in both improving the enhanced epistemic trustworthiness of AI systems or human experts and providing indications of that trustworthiness. For instance, when a certification is required to perform a certain activity (e.g. practicing medicine, using an instrument or being competitive in the AI marketplace) the

certification first provides the trustee with motivation to follow certain rules and complete certain sorts of training (in order to get the certification) and then it functions to signal that the certified entity has met those requirements. Similarly, epistemic safety in the context of AI production might be both improved and communicated via the implementation of public engagement programs by AI developers. Efforts to involve public interest groups in, for example, the conceptualization of guiding principles or discussions about their hopes and fears regarding proposed AI applications helps the AI producer learn how to align its work with public interests and values and simultaneously communicates to the public that the AI producer takes user interests seriously.

There is a challenge, however, that even when efforts are undertaken to communicate enhanced epistemic trustworthiness to users, that those efforts will not necessarily help users form well-founded beliefs about the epistemological well-functioning of a network of systemic factors. We run into similar challenges to those I discussed in Chapter 2 with respect to novice appraisal of expert explanations and in Chapter 3 with respect to internal justification for idiosyncratic trust in experts and AI systems; it is unreasonable to expect that the typical AI user will have the time, effort, and familiarity with tech policy or the technology in question to determine for herself that existing systemic factors do indeed constitute an epistemically well-functioning system, let alone to identify the relevant systemic factors. It is for this reason that things like credentials and certifications are used as second-order proxies to flag that systemic factors are in place. It does not necessarily follow, however, that those systemic factors are well implemented or effective at influencing the quality of an AI product or how it is employed. For example, with a quick Web search, anyone can see that Microsoft's Office for Responsible AI (ORA), Aether Committee, and Champs program do exist, and one can get a general sense for what they are supposed to achieve, but it is unlikely that the lay reader will be able to appraise for herself whether those programs are in fact likely to have a positive impact on the quality of Microsoft's AI products. Indeed, even I, who has given these issues considerable thought, cannot say for certain to what extent Microsoft's efforts might be another instance of ethics washing.

As such, systemic trust is almost always heavily based in assumption – assumptions that effective systemic factors are in place, that those factors were well thought out when implemented, and that someone else (or a collection of someones) maintains the epistemological well-functioning of the system. These assumptions are not, however, uninformed. When we think of cases where systemic trust seems strongest as in medical settings, our reasons for trusting the medical practitioner are largely based in assumptions about medical systems that regulate physician conduct. Those assumptions are in turn based in an awareness that the medical field is old, that principles or medical ethics and strategies for their implementation have been developed and iterated upon over hundreds of years, that many organizations exist with the express purpose of maintaining safety and quality of care (e.g. CDC, FDA, NIH, NHS, and

BMA), and that issues related to medical safety and efficacy are taken seriously as they are debated, discussed, and meticulously scrutinized in academic settings and on the public stage. Internal justification for systemic trust is stronger the more difficult it seems it would be for mistakes and misconduct to go unnoticed and unaddressed.

What does this all mean for internal justification for systemic trust in AI? Unlike medicine, AI research and development is new, and as I have demonstrated, few factors are currently in place to form the basis of epistemologically well-functioning systems around AI production and employment, though on a more positive note, clear efforts are underway to change this state of affairs. For example, the European Commission's extensive AI policy recommendation may be implemented as early as Summer 2022, and if the GDPR is anything to go by, many other countries are likely to follow suit. However, even if and when such policies are in place, systemic trust will not immediately follow. Engendering enhanced systemic trust takes time, and that time must be characterized by ongoing and publicly visible appraisals of implemented measures, discussions about the principles on which those measures are based, and continued efforts to show that concerns about AI ethics and safety are taken seriously by policymakers, and AI producers, and employers alike. The very fact that debates around AI principles, best practice, and safe and socially responsible employment are ongoing plays some role in internal justification for systemic trust insofar as they imply that key actors are thinking seriously about establishing well-functioning systems of AI production and employment.

6

Conclusion:
Epistemic security in a technologically advanced world

In this dissertation I set out to investigate how users of AI-enabled expert systems can make epistemically well-informed decisions about believing AI system outputs (predictions, recommendations, etc.). The overarching goal of this investigation has been to gain a better understanding of how the deployment of AI-enabled expert systems into roles traditionally filled by human experts impact epistemic security. Epistemic security describes the state in which a person is able to consistently access and/or identify epistemically well-founded information or epistemically trustworthy information sources (Seger et al., 2020). However, AI-enabled expert systems are but one kind of AI application, and AI applications are but one class of technology that potentially pose challenges to epistemic security. Therefore, to conclude this dissertation, I take a step back to briefly summarize and reframe my project against the much broader topic of epistemic security in a technologically advanced world.

Epistemic security is important for a variety of reasons. From the perspective of an individual, epistemic security is key to making informed decisions about what to believe and, in turn, how to act in order to achieve a goal. For example, if a person believes the AI diagnostic app on her phone suggesting that the weird mole on her foot is not cancerous, she is not likely to go through the inconvenience of booking an appointment to have the mole biopsied so that a pathologist (human or

AI) can confirm. From a societal perspective, epistemic security is essential to coordinating the kind of collaborative action necessary to respond to complex challenges and crises like climate change and global pandemics. When various actors receive conflicting decision-guiding information, it becomes much more difficult to coordinate a unified, timely, and effective response to the challenge at hand.

In our modern, technologically advanced world, epistemic security and threats to it cannot be discussed without paying close attention to the impact of information producing and mediating technologies on how people acquire and process information. There are a variety of ways technologies impact information availability and acquisition. To start, instruments act as epistemic enhancers enabling us to make observations, remember information, or reason to conclusions that would otherwise be outside the reach of our natural powers of perception and rational capacities (Humphreys, 2004). A microscope, for example, brings into view objects too small for the human eye to perceive, and AI systems statistically process vast sets of data in order to identify patterns that would otherwise go unnoticed. Other technologies like video and audio recording and digital storage augment human memory, while online search and search recommendation systems help us locate information relevant to our queries among the masses of data we are now able to accumulate.

Modern technology can promote epistemic security by helping us access, remember, and appraise information, but it can also introduce complications. A scratched microscope slide distorts the image produced, and a biased training data set negatively impacts the quality of an AI system's predictions. Photoshop and deepfake capabilities alter or create recorded memories (Rini, 2020; Chesney & Citron, 2019), and online search recommendation algorithms influence the kinds of questions people ask and, in turn, the answers they receive (Miller & Record, 2017).[1] In mediating how we remember and access information, technological instruments introduce opportunity for error in our belief formation processes. As such, forming justified belief in the things we hear, read, and see in a technologically enabled age often requires holding some belief about the reliability or trustworthiness of the technologies producing or mediating the information we receive. The impacts

[1] Autocompleting search recommendation systems have been found to increase the speed of online search inquiries and the quality of online search results overall (Kato et al., 2013; Ward et al., 2012; White & Marchionini, 2007). However, Boaz Miller and Isaac Record warn that auto recommendation can also epistemically harm "the *searcher*... when she develops false, biased, or skewed beliefs" based on the recommended completions, and can epistemically harm "*society as a whole*... when prejudicial, biased, or false beliefs about members of different groups in society lead to the incorrect assessment of the trustworthiness of informants who belong to these groups and, consequently, to blocking the circulation of critical ideas in society" (Miller & Record, 2017: 1951). For example, when I type "Women should..." into the Google search bar, the first three autocomplete recommendations are as follows: (1) "Women should...dress modestly," (2) "Women should...be financially independent," and (3) "Women should...be prim and proper" (Google search as of December 12. 2021). While many people would initially disregard the associations suggested by autocompletions 1 and 3, Miller and Record (2017: 1949) warn that "involuntary exposure to certain autosuggestions [like 1 and 3] may reinforce unwanted beliefs" because people are bad at recognizing and overcoming their implicit biases. Furthermore, "autosuggestions interactively affect a user's inquiry, leading to paths she might not have pursued otherwise". For instance, a person might abandon the search altogether if she believes the autosuggestion provided a satisfactory answer. In these ways autocompletion algorithms can influence how people acquire information and what beliefs they form about the topics they search.

of modern information producing and mediating technologies on epistemic security is a rich field of investigation. In this dissertation I took one small, though none-the-less important, bite. I focused on a particular kind of technology – AI-enabled expert systems – and evaluated whether these technologies pose any new, or exacerbate any existing, challenges to information acquisition, appraisal, and/or belief formation.

I have focused on AI-enabled expert systems for a couple of reasons. First, they are becoming more ubiquitous, operating in roles traditionally filled by human experts. I began this dissertation by describing expert testimony as arguably the most important source of knowledge. Epistemology of testimony is the study of how people acquire knowledge from others, and the vast majority of what we think we know about the world we do learn from others. Testimony has enabled the accumulation of significant bodies of human knowledge over time, and reliance on testimony is necessary to executing large-scale projects and for responding to complex challenges and crises that require collaboration across time and areas of expertise. It is therefore key to the stability of society that we understand how people can justify their belief in the information they receive via testimony. I argue that in our modern technological age, understanding how people acquire knowledge from others must extend to understanding how people justify their beliefs in information acquired from information mediating and producing technologies as well. So where AI systems start producing outputs in place of human expert testimony, we need to pay attention.

Second, the challenges that AI systems pose to epistemic security seem more complicated than those posed by prototypical instruments. Where there is concern that an instrument like a microscope or a thermometer is prone to error, one can either appraise its internal mechanisms for the error source herself, or consult another human who has the requisite expertise and familiarity with the instrument to do so. With an AI system, however, there is an issue of the system's epistemic opacity; AI systems are too complex for any human, lay user or AI developer, to trace in detail how system outputs are derived from inputs. It is not possible to just peek inside an AI system like one would inspect any other mechanical instrument in order to pinpoint existing or potential sources of error in the system's output derivation processes. As such, looking into AI-enabled expert systems has allowed me to ask whether these systems really do pose any new or more significant challenges to epistemic security, or if this concern is just part of the hype around a novel technology.

Throughout this dissertation I have answered in the negative: AI-enabled expert systems do not pose new challenges to epistemic security, nor do they present hurdles that are much higher than those hurdles grappled with in an AI-free world. The challenges are significant, but not new. My position, however, is not based on a comparison between AI systems and prototypical instruments, but is based on a comparison between AI systems and human experts – a comparison which I have called *adopting the testimonial stance*. By adopting the testimonial stance, one takes seriously any similarities between AI systems and human testimonial speakers. I note that both AI systems and

human experts are epistemically opaque: it is difficult to trace, in detail, the processes by which each derives its conclusions. Furthermore, both AI-user and expert-novice relationships are plagued by a version of Meno's paradox: in virtue of their epistemically disadvantaged position relative to the information source, a novice or lay user is unable to independently verify the content of the expert's or AI's claims or recommendations. Indeed, if she did have both the cognitive capacity and the time to do so, there would have been no reason to seek expert or AI advice in the first place. Consequently, it is not clear how a novice or lay user can gain internal epistemic justification for her belief in expert claims or AI outputs.

I have argued that in order to adopt the testimonial stance, one need not first defend AI-enabled expert systems as sources of knowledge from expert testimony as opposed to knowledge from observation, memory, or reason mediated by instruments. Very broadly, epistemology of testimony is concerned with the acquisition of knowledge from others' claims. While testimonial speakers are typically assumed to be human, I have argued that, when it is useful to do so, the classification may be extended to information producing artifacts by thinking of them *as if* testimonial speakers. The question now is whether it has been useful for me to adopt the testimonial stance toward AI-enabled expert systems. Has the comparison between human experts and AI-enabled expert systems guided a fruitful investigation into how users of these systems might acquire a degree of internal epistemic justification for their belief in system outputs? Indeed it has, and not just for shedding light on how users of AI technologies can justify their belief in AI outputs, but for better understanding the nature of justified belief in expert claims as well.

First in Chapter 1, the testimonial stance helped in establishing a philosophical approach to thinking about AI generated explanations. Like human expert-to-novice explanations, AI explanations are post hoc and partial communications of the processes by which an AI system derived an output (by which an expert reasoned to her conclusion). These post hoc accounts are reduced in detail and often abstracted in reflectivity for the sake of recipient understanding. This notion of explanation as post hoc and partial accounts of "reasoning" does not, however, align well with the paradigm notion of explanation as understood by most philosophers of scientific explanation. As such, I cautioned that one can easily get lost in the explanations literature while investigating if and how expert and AI explanations provide recipients with information they can use to help justify their belief in AI outputs and expert claims. I then recommended consulting the literature on the contexts of discovery and justification in scientific research to help guide the discussion. The discovery-justification literature is not often connected to issues of explanation, though I argue it should be. The literature is motivated by the disconnect that exists between what scientists do to derive their findings in the context of discovery and what they report of those processes (that is, how they explain their findings) in the context of justification. The discovery-justification distinction keeps in focus the question we need to ask when investigating the justificatory value of AI and expert explanations: How do abstractions in

explanation detail and reflectivity from the original account impact a recipient's ability to use that information to appraise the veracity or well-foundedness of the expert's claim or an AI system's output? In short, I described how abstractions generally contribute to an explanation's justificatory value by making explanations understandable (explanation understandability is a prerequisite to justificatory value) and detract from its justificatory value by reducing the amount of information contained in the explanation for recipients to analyze.

In Chapter 2, I finally tackled the looming question: what is the justificatory value of an AI-to-user or Expert-to-novice explanation? Building on a coherence theory of justification, I demonstrated how the greater the epistemic divide between explainer and explainee, the less justificatorally valuable an explanation can be. This is bad news for explanations offered to lay AI users and the more naïve novice. Due to the wide epistemic divide between information source and sink, there is only so much the recipient can do with the information contained in an explanation to help epistemically justify her belief in an expert claim or AI output.

Please note: I am not arguing that explanations are completely useless. As discussed at the end of Chapter 2, AI and expert explanations have various important functions in addition to enabling epistemically justified belief. For instance, explanations provide opportunity for recipients to ask questions; in certain cases, like when receiving a medical diagnosis, an explanation might help a person psychologically and emotionally prepare for the risks and consequences of treatment; or an explanation might provide indications of the values and priorities underpinning a reasoning process and whether they align with the recipient's own (Manson, 2010). Explanations could also be used to set realistic expectations for the kinds of questions an AI system is designed to answer and the kinds of answers a system might provide. In this way AI explainability can be seen as facilitating more natural human-computer interactions and thereby improving human-AI team performance, as when a physician consults AI diagnostic and treatment planning tools to help treat a patient (Vorvoreanu & Walker, 2022). So, to be clear, it has not been my position that AI explainability work should be abandoned. These other functions should not be dismissed. Indeed, explanations may play a role in helping the recipient appraise the alignment of a system's outputs with her own interests and values which, as discussed in Chapter 5, is necessary for enhanced epistemic trust. However, the basic epistemic concern upon which enhanced epistemic trust builds still remains: for lay users and novices, explanations provide little epistemic basis for believing the outputs and claims of experts and AI systems alike. So, when it comes to basic epistemic justification, looking to AI explainability to aid users in validating the content of AI outputs for themselves is barking up the wrong tree. Thus ends Part 1.

In Part 2, I turned my discussion fully to the topic of trust. Due to the limited justificatory value of explanations in epistemically imbalanced relationships, I argued that epistemic trust in the information source must be the principal foundation for epistemic justification in expert and AI claims

and outputs. Many people will balk at the idea of using trust terminology with respect to AI systems and other artifacts, but in adopting the testimonial stance I bypassed this theoretical quibble in order to benefit from the practical discussions that existing discourse on trust in expertise have to offer. For those still doubtful, I interjected with my own account of trust as distinct from reliance in Chapter 4. According to my account, trust and trustworthiness may be used to refer to humans and artifacts alike. However, to avoid getting sidetracked in these conceptual weeds, I first set the agenda for Part 2 by launching a discussion about systemic trust in human experts and AI in Chapter 3.

I argued in Chapter 3 that systemic trust, as opposed to idiosyncratic trust, is the primary foundation for epistemic trust in both human experts and AI systems. Systemic trust is the kind of trust a person holds in an external entity not because of trustee's idiosyncratic characteristics (e.g. an expert's skill or sincerity, or an AI system's safety features or training data sets) but because she is aware that there exists a larger network of systemic factors in which the AI/expert is embedded which serve to maintain minimum performance standards. Systemic factors include both intrinsic factors (like community values and norms) and extrinsic factors (like explicit rules and regulations which) which regulate both the contexts of AI/expert employment and production.

Overall, it seems plausible that it is easier for lay users and novices to gain some internal epistemic justification for systemic trust in AI and experts than to gain epistemic justification for belief in AI outputs and expert claims via explanations. If, based on a lay user's/novice's awareness of the systemic factors at play, it seems that it would be quite difficult for errors or misconduct to go unnoticed and unaddressed, then the user/novice would have reason to believe that an expert or AI system embedded in that network of systemic factors is a trustworthy source of information who is likely to produce true or epistemically well-founded outputs. Accordingly, the user/novice would also hold a degree of epistemic justification for believing the outputs/claims of the AI/expert. However, as I explained in Chapter 5, the relative ease of gaining information about systemic factors underpinning epistemically well-functioning systems of AI employment and production may be outweighed by the concern that, as of yet, there is little information to be gained. At this moment in time, there is little foundation for systemic trust in AI either in terms of widespread AI community commitment to common ethical and epistemic standards or the implementation of effective AI governance structures to enforce those standards, though I suggest we may be optimistic about the future.

So, to summarize, in Part 1 I determined that while in theory it is possible for expert and AI explanations to provide recipients with information they can use to help epistemically justify their belief in expert claims and AI outputs, in practice the justificatory value is often quite low due the epistemically imbalanced nature of AI-user and expert-novice relationships. In Part 2 I demonstrated that it is more likely that novices and lay users will be able to acquire a degree of epistemic justification for believing the claims of expert and AI-enabled expert systems via systemic trust held in the expert(-system). Accordingly, I have concluded that it is key to the epistemic security of

societies in which people depend on expert testimony and AI-enabled technologies, that systems of AI and expert production and employment strive to be epistemically well-functioning and to demonstrate that well-functioning to the public. I define epistemically well-functioning systems as those in which various intrinsic and extrinsic systemic factors effectively influence individual expert/AI performance (the sociological premise is satisfied), and in which those systemic factors promote epistemic and ethical standards which, when satisfied, are conducive to expert trustworthiness (the enhanced epistemological premise is satisfied).

So now I return to the question I posed a few pages back: has adopting the testimonial stance toward AI-enabled expert systems guided a fruitful investigation into how users of these systems might acquire a degree of internal epistemic justification for their belief in system outputs? It clearly has. Overall, adopting the testimonial stance toward AI-enabled expert systems has provided a unique and productive framework to guide thought and discussion about the challenges to justified belief in AI outputs. Viewing AI-enabled expert systems as akin to human experts has established the epistemically imbalanced nature of AI-user and expert-novice relationships as the center of analysis as opposed to the artificiality of machine intelligence. Some may worry that comparing AI systems to human experts unnecessarily anthropomorphizes and sensationalizes AI technologies. For example, frequent comparison between artificial intelligence and human intelligence can lead to exaggerated hopes and fears for AI technology which undermine accurate appraisals of system trustworthiness (Bryson & Kime, 2011; Cave, Coughlan & Dihal, 2019; Cave & Dihal, 2019). This, in turn, may lead to unwarranted reliance on an unsafe technology, or distrust in a technology which might otherwise benefit humanity greatly. However, I have demonstrated, to the contrary, that the move to compare AI systems to human experts via the testimonial stance guards against unwarranted catastrophizing. The challenges to justified belief in the outputs of AI-enabled expert systems are significant, but they are not new, and in both the case of AI-enabled expert systems and human experts, the key to epistemic security is building systemic trust; due to the epistemically imbalanced nature of novice-expert and user-AI relationships, novice/user sensitivity to the veracity of expert claims and AI outputs and to the idiosyncratic trustworthiness of individual human experts and AI systems is often low. Therefore, epistemic security must involve building, and indicating the presence of, epistemically safe networks of systemic factors around experts and AI. The idea is that in such networks of systemic factors, it would be difficult to go wrong in believing the outputs of AIs/experts embedded therein.

To conclude this dissertation, I would like to return to where this chapter began, with the observation that AI-enabled expert systems constitute only a narrow subset of modern information producing and mediating technologies which might pose challenges to epistemic security. In limiting my discussion to AI-enabled expert systems I have only touched the tip of the iceberg in investigating technologically exacerbated threats (or, conversely, technologically enabled benefits) to epistemic security. To begin, there are other AI applications which not only produce outputs but can also be used

to manipulate how people perceive and form beliefs about those outputs by hijacking the naturally ingrained heuristics that people use to evaluate human speaker trustworthiness. For example, Deepfake audio and video can identify and mimic speech patterns, vocal tones, and facial expression that users respond to positively in order to attract user trust irrespective of the content being communicated (Rini, 2020; Chesney & Citron, 2019), and web bots can be used to make minority extremist opinions seem much more widely accepted than they actually are (Chessen, 2017). AI powered content targeting applications can also be used to identify individuals most susceptible to different kinds of advertising content. The UK based consulting firm Cambridge Analytica, for instance, claims to have three to five thousand data points on more than 230 million US adults drawn from Facebook user data (Funk, 2016; Cadwalladr & Graham-Harrison, 2018). The firm uses the data to build psychological profiles of each individual so as to identify those most easily swayed in opinion and to target them with the most effective campaign material in order to influence their belief and opinion formation processes.

Technological threats to epistemic security also expand beyond the impacts of AI-enabled technologies. For instance, global connectivity enabled by the internet combined with the ease of communication and ability to remain anonymous online makes it easier than ever for people to either intentionally or accidentally spread false and misleading information to a wide audience. Furthermore, the sheer abundance of information that is easily accessible and actively pushed at people online strains the limited human capacity for attention (Roetzl, 2019). This has resulted in the development of a fierce attention economy in which information producers and distributors must compete for real estate in the forefront of the human mind by constantly developing new attention-grabbing strategies like designing eye-catching user interfaces or engineering appeals to user emotion (Simon, 1971). The problem, however, is that these strategies are truth neutral in that they will draw user attention regardless of the quality of the information being promoted (Seger et al., 2020: 26-29).

While my focus in this dissertation has been on the epistemic security of user interactions with AI-enabled expert systems, the strategies I have discussed for promoting epistemic security are not, however, limited to addressing challenges posed by AI. Like AI-enabled expert systems, other information producing and mediating technologies are not inherently evil. They can be of great epistemic benefit in enabling easy access to quality information and are essential to the coordination of collaborative efforts and collective action on a global scale. Problems emerge when and where these technologies are misused. Therefore, the best approach to promoting epistemic security in technologically advanced societies is not to completely do away with the technologies posing epistemic threats, but rather to attend to how the technologies are used and regulated. Systemic trust is still key.

Networks of intrinsic and extrinsic systemic factors are needed to encourage a culture of responsibility among information technology developers and employers, to enforce restrictions that

prevent the deployment of epistemically unsafe and/or untested technologies, and to otherwise guard against the intentional or accidental misuse of information producing and medicating technologies. In the original epistemic security report, my colleagues and I put forward a few suggestions for building more epistemically secure information ecosystems in our technologically advanced world (Seger et al., 2020: 41-45). For example, we suggest the establishment of financial penalties for the knowing dissemination of unsupported, fabricated, or false information or for utilizing information technologies to aid in the production and dissemination of information without satisfying minimal "epistemic responsibility" standards to guard against the perpetuation and magnification of misinformation. In a different vein, we also recommend building capacity within academia and/or publicly funded research institutes to identify and forecast current and emerging technological threats to epistemic security, to monitor for changes in our epistemic ecosystems, and to advise on technologically enabled communication strategies for facilitating timely and well-coordinate crisis response.

None of this is to say that pursuing epistemic security is a straightforward or easy process. To the contrary, the sociotechnical epistemic processes by which information is produced, disseminated, acquired, and evaluated in heterogeneous and technologically advanced societies are complex and heavily intertwined. This means that there are many points of epistemic vulnerability within social epistemic systems and that narrowly targeted fixes to address threats to epistemic security can easily have unintended second-, third-, and higher-order consequences. Accordingly, my colleagues and I have also argued that effectively identifying epistemic threats and vulnerabilities and scanning for potential unintended consequences will require eliciting the input of a diverse array of experts who work in and around the epistemic security space (Seger et al., 2020: 44-45). For example, responsible journalists and journalism agencies engage in internal fact checking procedures to counter the spread of misinformation, and external fact checking organizations within universities and independent research centers engage in activities to encourage fact-based public discourse and promote accurate beliefs among the public and policy makers (Graves & Amazeen, 2019). Psychologists investigate vulnerabilities in the processes by which individuals choose to consume information and form beliefs (Kahneman, 2011; Klayman, 1995) and epistemologists study conditions that influence the justification of belief. Technologists can further comment on what interventions are technologically feasible. Furthermore, pursuing epistemic security is not just about eliciting input from diverse experts. It is also important to encourage cross-disciplinary thinking strategies. In short, this means not immediately excluding a line of reasoning because it crosses a theoretical line held by one discipline or another (like thinking about AI-enabled expert systems as if they are sources of testimony). It means drawing from a variety of viewpoints to find tools to guide our investigations where they happen to exist and exploring their combination where helpful.

BIBLIOGRAPHY

Achinstein, P. (1983). *The Nature of Explanation*, Oxford: Oxford University Press.

Achinstein, P. (2010). *Evidence, Explanation, and Realism: Essays in Philosophy of Science*. Oxford: Oxford University Press.

Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access, 6*, 52138-52160. doi:10.1109/ACCESS.2018.2870052

Anderson, E. (2006). The epistemology of democracy. *Episteme, 3*(1), 9-23.

Anderson, E. (2011). Democracy, public policy, and lay assessments of scientific testimony. *Episteme, 8*(2), 144-164. doi:10.3366/epi.2011.0013

Anderson, E. (2021). Epistemic Bubbles and Authoritarian Politics. In M. Hannon & E. Edenberg (Eds.), *Political Epistemology* (pp. 11-30). Oxford: Oxford University Press.

Anderson, M. (2021). Deepfaked Voice Enabled $35 Million Bank Heist in 2020. *Unite.AI*. Retrieved from https://www.unite.ai/deepfaked-voice-enabled-35-million-bank-heist-in-2020/

Angwin, J., Larson, J., Mattu, S., & Lauren Kirchner. (2016). Machine Bias. *ProPublica*. Retrieved from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Anik, L., et al. (2013). Prosocial Bonuses Increase Employee Satisfaction and Team Performance, *PLoS ONE, 8*(9). https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0075509 (link as of 26/11/20)

Ashurst, C., Anderljung, Markus, Prunkl, C., Leike, J., Gal, Y., Shevlane, T. & Dafoe, A. (2020). *A Guide to Writing the NeurIPS Impact Statement.* Retrieved from https://medium.com/@GovAI/a-guide-to-writing-the-neurips-impact-statement-4293b723f832#005f

Audi, R. (1998). *Epistemology: A Contemporary Introduction to the Theory of Knowledge*. New York: Routledge.

Ayer, A. J. (1956). *The Problem of Knowledge. London: Macmillan.*

Baehrens, D., Schroeter, T., Harmeling, S., Kawanabe, M., Hansen, K., & Müller, K. R. (2010). How to explain individual classification decisions. *J. Mach. Learn. Res., 11*(6), 1803–1831.

Baier, A. C. (1986), Trust and Antitrust, *Ethics* 96.2: 231– 60.

Baier, A. C. (1994). *Moral Prejudices: Essays on Ethics.* Cambridge, MA: Harvard University Press.

Baum, S. D. (2017). On the promotion of safe and socially beneficial artificial intelligence. *AI & Society, 32*, 543-551. doi:10.1007/s00146-016-0677-0

Beauchamp T. & Childress J. (1979). *Principles of Biomedical Ethics*. New York: Oxford University Press.

Bechtel, W. (2011). Mechanism and Biological Explanation. *Philosophy of Science, 78*(4), 533-57.

Becker, K. (2009). Margins for Error and Sensitivity: What Nozick Might Have Said. *Acta Analytica 24*(1), 17 –31.

Beecher H. K. (1966). Ethics and Clinical Research. *N. Engl. J. Med,* 274(24), 1354–60.

Benjamin, R. (2019). Assessing risk, automating racism: A health care algorithm reflects underlying racial bias in society. *Science, 366*(6464), 421-422. doi:10.1126/science.aaz3873

Bien, J. and Tibshirani, R. (2011). Prototype selection for interpretable classification. *Ann. Appl. Statist*., *5*(4), 2403–2424.

Bird, A. (2014). When is there a group that knows? Distributed cognition, scientific knowledge, and the social epistemic subject. In J. Lackey (Ed.), *Essays in Collective Epistemology* (pp. 42-63).

Bishop, M. A., & Trout, J. D. (2002). 50 years of successful predictive modeling should be enough: lessons for philosophy of science. *Philosophy of Science, 69*(23), S197-S208.

Bishop, M. A., & Trout, J. D. (2005). The amazing success of statistical prediction rules. In M. A. Bishop & J. D. Trout (Eds.), *Epistemology and the Psychology of Human Judgment* (pp. 24-53). Oxford: Oxford University Press.

Black, T. (2008). Defending a Sensitive Neo-Moorean Invariantism. in V. F. Hendricks and D. H. Pritchard, (eds.), *New Waves in Epistemology* (pp. 8-27): Basingstoke: Palgrave Macmillan.

Black, T. and Murphy, P. (2007). In Defense of Sensitivity. *Synthese 154*(1): 53 –71.

Boden, M. (2014). GOFAI. In K. Frankish & W. Ramsey (Eds), *The Cambridge Handbook of Artificial Intelligence* (pp. 89-107). Cambridge: Cambridge University Press. doi:10.1017/CB09781139046855.007

Boland, P. J. (1989). Majority Systems and the Condorcet Jury Theorem. *Journal of the Royal Statistical Society*, Series D (The Statistician), 38, 181–9.

Bolinska, A. (2013). Epistemic representation, informativeness and the aim of faithful representation. *Sythese, 190*(2), 219-234. doi:10.1007/s11229-012-0143-6

Bokulich, A. (2009). Explanatory Fictions. In M. Suárez (Ed.), *Fictions in Science: Philosophical Essays on Modeling and Idealization* (pp. 91-109): Routledge.

Bokulich, A. (2011). How Scientific Models Can Explain. *Synthese*, *180*(1), 33–45. doi:10.1007/s11229-009-9565-1

Bokulich, A. (2016). Fiction as a vehicle for truth: Moving Beyond the Ontic Conception. *The Monist, 99*(3), 260-279.

BonJour, L. (1985). *The Structure of Empirical Knowledge*. Cambridge: Harvard University Press.

Bouchard, F. (2016). The roles of institutional trust and distrust in grounding rational deference to scientific expertise. *Perspectives on Science, 24*(5), 582-608. doi:10.1162/POSC_a_00224

Brennan, T., Dieterich, W. & Ehret, B. (2009). Evaluating the Predictive Validity of the COMPAS Risk and Needs Assessment System. *Criminal Justice and Behavior, 36*, 21–40.

Brothers, K. B., Rivera, S. M., Cadigan, R. J., Sharp, R. R., & Goldenberg, A. J. (2019). A Belmont Reboot: Building a Normative Foundation for Human Research in the 21st Century. *The Journal of law, medicine & ethics: a journal of the American Society of Law, Medicine & Ethics*, *47*(1), 165–172. doi:10.1177/1073110519840497

Brundage, M., Avin, S., Wang, J., Belfield, H., Kreuger, G., Hadfield, G., ... Seger, E., ... Anderljung, M. (2020). Toward Trustworthy AI Development: Mechanism for supporting verifiable claims. arXiv:2004.07213

Bryson, J. J. (2010). Robots Should Be Slaves. In Y. Wilks (Ed.), *Close Engagements with Artificial Companions: Key social, psychological, ethical and design issues* (pp. 63-64): John Benjamins.

Bryson, J. J. (2018a). Patiency is not a virtue: the design of intelligent systems and systems of ethics. *Ethics Inf Technol 20,* 15-26. https://doi.org/10/1007/s1067-018-9448-6

Bryson, J. J. (2018b). AI & global governance: No one should trust AI. United Nations University Centre for Policy Research Article. URL: https://cpr.unu.edu/publications/articles/ai-global-governance-no-one-should-trust-ai.html

Bryson, J.J., Kime, P. P. (2011). Just an artifact: Why machines are perceived as moral agents. In *Proceeding of the 22nd International Joint Conference on Artificial Intelligence* (pp.1641-1646), Barcelona: Morgan Kaufmann. doi:10.5591/978-1-57735-516-8/IJCAI11-276

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society, 3*(1), 1-12. doi:10.1177/2053951715622512

Cadwalladr, C., & Graham-Harrison, E. (2018). Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach. *The Guardian*. https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election (accessed April 22, 2021)

Calo, R. (2017). Artificial intelligence policy: a primer and roadmap. *UC Davis Law Review, 51*, 399–436.

Cariani, F. (2011). Judgment Aggregation. *Philosophy Compass, 6*(1), 22-32. doi:10.1111/j.1747-9991.2010.00366.x

Carr, N. (2015). *The glass cage: Who needs humans anyway?* UK: Vintage Publishing.

Carson, T. L. (2010). *Lying and Deception: Theory and Practice.* Oxford University Press: Oxford.

Cartwright, N. (1983). *How the Laws of Physics Lie*. Oxford: Oxford University Press. doi:10.1093/0198247044.001.0001

Caruana, R., Lou, Y., Gerhke, J., Koch, P., Sturm, M. & Elhadad, N. (2015). Intelligible Models for Healthcare: Predicting Pneumonia Risk and Hospital 30-Day Readmission. in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York: Association for Computing Machinery.1721-30. doi:10.1145/2783258.2788613

Cave, S., Coughlan, K., & Dihal, K. (2019). *Scary robots: Examining public response to AI.* Paper presented at the 2019 AAAI/ACM Conference on AI, Ethics, and Society, Honolulu, HI.

Cave, S., Dihal, K. (2019). Hopes and fears for intelligent machines in fiction and reality. *Nat Mach Intell 1***,** 74–78. doi:10.1038/s42256-019-0020-9

Cave, S., Whittlestone, J., O hEigheartaigh, S. & Calvo, R. A. (2021). *Using AI ethically to tackle covid-19. BMJ, 372*(364). doi:10.1136/bmj.n364

Chesney, R., & Citron, D. (2019). Deepfakes and the new disinformation war: the coming age of post-truth geopolitics. *Foreign Affairs, 98*(1), 147-155.

Chessen, M. (2017). The MADCOM future: How artificial intelligence will enhance computational propaganda, reprogram human culture, and threaten democracy… and what can be done about it. Atlantic Council, Dinu Patriciu Eurasia Center, and Brent Scowcroft Center on International Security. Retrieved from https://www.atlanticcouncil.org/wpcontent/uploads/2017/09/The_MADCOM_Future_RW_0926.pdf

Chisholm, R. M. (1957). *Perceiving: A Philosophical Study*. Ithaca, New York: Cornell University Press.

Clouser, K. D., & Gert, B. (1990). A critique of principlism. *The Journal of Medicine and Philosophy, 15*, 219-236.

Coady, C. A. J. (1992). *Testimony. A Philosophical Study*. Oxford: Oxford University Press.

Coeckelbergh, M. (2012). Can we trust robots? *Ethics and information technology, 14*, 53-60. doi:10.1007/s10676-011-9279-1

Coeckelbergh, M. (2019). Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability. *Science and Engineering Ethics*, 1-18. doi:https://doi.org/10.1007/s11948-019-00146-8

Coleman, J. S. (1994). *Foundations of Social Theory*. Cambridge, MA: Harvard University Press.

Collett, C. & Dillon, S. (2019). AI and Gender: Four Proposals for Future Research. Cambridge: The Leverhulme Centre for the Future of Intelligence. http://lcfi.ac.uk/media/uploads/files/AI_and_Gender_4_Proposals_for_Future_Research_210619_p8qAu8L.pdf

Collins, H., & Evans, R. (2007). *Rethinking Expertise*. Chicago, IL: University of Chicago Press.

Collins, H., & Weinel, M. (2011). Transmuted Expertise: How Technical Non-Experts Can Assess Experts and Expertise. *Argumentation, 25*, 401-413. doi:10.1007/s10503-011-9217-8

Comesaña, J. & Klein, P. (2019). Skepticism. In E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy.* https://plato.stanford.edu/entries/skepticism/#KnowJustSkep (accessed May 15, 2022).

Conee, E., & Feldman, R. (2004). *Evidentialism: Essays in Epistemology.* Oxford: Clarendon Press. Steup, M. (1996). An Introduction to Contemporary Epistemology. Upper Saddle River, NJ: Prentice-Hall.

Cooter, R. D. (2000). Three effects of social norms on law: expression, deterrence, and internalization. *Oregon Law Rev* 79:1–22

Craver, C. F. (2014). The Ontic Conception of Scientific Explanation, in Andreas Hutteman and Marie Kaiser (eds.), *Explanation in the Biological and Historical Sciences,* Springer.

Craver, C. F., & Darden, L. (2013). *In Search of Mechanisms: Discoveries across the Life Sciences.* The University of Chicago Press.

Cullen, Z. B. & Perez-Truglia, R. (2020). The Old Boys' Club: Schmoozing and the Gender Gap. *National Bureau of Economic Research*. doi:10.3386/w26530.

Dabkowski, P. & Gal, Y. (2017). Real time image saliency for black box classifiers. *Proc. Adv. Neural Inf. Process. Syst*, 6970–6979.

Davis, R. B. (1995). The principlism debate: A critical overview. *The Journal of Medicine and Philosophy, 20*, 85-105.

DeepMind Health Independent Review Panel Annual Report. (2017). Retrieved from https://kstatic.googleusercontent.com/files/7e0b35e4cb6ccb750cba03fb160a69cc4f24456358042b8313b88943c49dfbce46037e9c89fad32fae986bd08a84e90c792656e0208d1276f1db895dcb42386b (accessed April 7, 2020)

DeepMind Health Independent Review Panel Annual Report. (2018). Retrieved from https://www.trusttech.cam.ac.uk/files/2018_deepmind_health_independent_review_annual_report.pdf (accessed April 7, 2020)

de Regt, H. W. (2009a). The epistemic value of understanding. *Philosophy of Science, 76*(5), 585-597.

de Regt, H. W. (2009b). Understanding and scientific explanation. In H. W. de Regt, S. Leonelli, & K. Eigner (Eds.), Scientific understanding: Philosophical perspectives (pp. 21–42). Pittsburgh: University of Pittsburgh Press.

de Regt, H. W., & Gijsbers, V. (2016). How false theories can yield genuine understanding. In S. R. Grimm, C. Baumberger, & S. Ammon (Eds.), *Explaining Understanding: New Perspectives from Epistemology and Philosophy of Science.* (pp. 50-75). New York: Routledge.

de Regt, H. W., Leonelli, S. & Eigner, K. (2009). *Scientific Understanding: Philosophical Perspectives*. Pittsburgh: University of Pittsburgh Press.

Dennett, D. C. (1987). *The Intentional Stance*. Cambridge: MIT Press.

Descartes, R. (1637). *Discours de la Méthode Pour bien conduire sa raison, et chercher la vérité dans les sciences*, (Discourse on the Method of Rightly Conducting the Reason and Seeking for Truth in the Sciences). Leiden: Jan Maire. [available online]

Dietrich, F., & Spiekermann, K. (2013). Independent Opinions? On the Causal Foundations of Belief Formation and Jury Theorems. *Mind, 122*(487), 655-685. doi:10.1093/mind/fzt074

Domenicucci, J., & Holton, R. (2017). Trust as a Two-Place Relation. In P. Faulkner & T. Simpson (Eds.), *The Philosophy of Trust*. Oxford, UK: Oxford University Press.

Douglas, H. (2000). Inductive Risk and Values in Science. *Philosophy of Science, 67*(4), 559-579.

Douglas, H. (2009). *Science, Policy and the Value-Free Ideal*. Pittsburgh: University of Pittsburgh Press.

Dreyfus, H. L., & Dreyfus, S. E. (1986). *Mind over machine: The power of human intuition and expertise in the era of the computer*. New York: Free Press.

Duval, A. (2019). Explainable Artificial Intelligence (XAI). *Unpublished Academic Report at University of Warwick*. doi:10.13140/rg.2.2.24722.09929

Elgin, M., & Sober, E. (2002). Cartwright on explanation and idealization. *Erkenntnis, 57*, 441–450.

Erasmus, A., Brunet, T.D.P. & Fisher, E. (2021). What is Interpretability?. *Philosophy & Technology, 34,* 833-862. doi:10.1007/s13347-020-00435-2

Esteva, A., Kuprel, B., Novoa, R. A., et al. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature, 542* (7639): 115-118.

Estlund, D. M. (1994). Opinion Leaders, Independence, and Condorcet's Jury Theorem. *Theory and Decision*, 36, 131–62.

European Commission (2020). *White paper on Artificial Intelligence - A European approach to excellence and trust*. Retrieved from: https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf

European Commission (2021). *Regulation of the European Parliament and of the Council: Laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts 2021* (Brussels). Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206

European Commission Independent High-Level Expert Group on Artificial Intelligence (2019). E*thics Guidelines for Trustworthy AI*. Retrieved from https://ec.europa.eu/futurium/en/ai-alliance-consultation.1.html

Faulkner, P. (2007). On Telling and Trusting. *Mind*. 116, 875–902.

Feldman, R. (2003). *Epistemology*. Upper Saddle River, New Jersey: Prentice Hall.

Fernandez-Loria, C., Provost, F., & Han, X. (2020). Explaining data-driven decisions made by AI systems: The counterfactual approach. doi:arXiv:2001.07417v3

Fidler, F. & Wilcox, J. (2021). Reproducibility of Scientific Results. In E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy.* https://plato.stanford.edu/archives/sum2021/entries/scientific-reproducibility/ (accessed April 22, 2022).

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... Vayena, E. (2018). AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines, 28*(4), 689-707. doi:10.1007/s11023-018-9482-5

Franklin, A., & Howson, C. (1998). Comment on 'The Structure of a Scientific Paper' by Frederick Suppe. *Philosophy of Science, 65,* 411–416.

Franklin, S. (2014). History, motivations, and core themes. In K. Frankish & W. Ramsey (Eds.), *The Cambridge Handbook of Artificial Intelligence* (pp. 15-33). Cambridge: Cambridge University Press. Doi: 10.1017/CBO9781139046855.003

Freiman, O., & Miller, B. (2020). Can artificial entities assert? In S. C. Goldberg (Ed.), *The Oxford Handbook of Assertion* (pp. 415-434): Oxford University Press.

Fricker, E. (2006). Testimony and epistemic autonomy. In J. Lackey & Ernest Sosa (Eds.), *The Epistemology of Testimony.* (pp. 225-250). Oxford: Oxford University Press.

Fricker, E. (2015). How to Make Invidious Distinctions amongst Reliable Testifiers. *Episteme, 12*(2), 173–202.

Fricker, M. (2012). Group testimony? The making of a collective good informant. *Philosophy and Phenomenological Research, 84*(2), 249-276.

Friedman, M. (1974). Explanation and Scientific Understanding, *Journal of Philosophy*, 71(1): 5–19. doi:10.2307/2024924

Funk, M. (2016). The Secret Agenda of a Facebook Quiz. *New York Times*, November 19, 2016, https://www.nytimes. com/2016/11/20/opinion/the-secret-agenda-of-a-facebookquiz.html (accessed April 6, 2022).

Furman, K. (2016). *AIDS denialism in South Africa: a case study in the rationality and ethics of science policy.* PhD thesis, London School of Economics and Political Science. Retrieved from http://etheses.lse.ac.uk/3443/

Galison, P. (2003). TheCollectiveAuthor. In P. Galison & Mario Biagioli (Eds.), *Scientific Authorship: Credit and Intellectual Property in Science*. (pp.327–55). New York: Routledge.

GDPR - General Data Protection Regulation (April 26, 2016). Accessed 22 November 2017 from http://ec.europa.eu/justice/data-protection/reform/files/regulation_oj_en.pdf

Gelfert, A. (2011). Expertise, argumentation, and the end of inquiry. *Argumentation, 25*(3), 297-312. doi:10.1007/s10503-011-9218-7

Gelfert, A. (2014). *A Critical Introduction to Testimony*. London: Bloomsbury.

Gettier, E. (1963). Is Justified True Belief Knowledge? *Analysis. 23*(6): 121 –23.

Gilbert, M. (2000). *Sociality and Responsibility: New essays in plural subject theory*. Lanham, MD: Rowman and Littlefield.

Gilbert, M. (2004). Collective epistemology. *Episteme*, *1,* 95–107.

Gillies, D. (2016). Evidence of mechanism in the evaluation of streptomycin and thalidomide. *Studies in History and Philosophy of Biological and Biomedical Sciences, 66*, 55-62. doi:https://doi.org/10.1016/j.shpsc.2017.06.003

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2019). Explaining Explanations: An Overview of Interpretability of Machine Learning. doi:arXiv:1806.00069v3

Glass, D. H. (2007). Coherence Measures and Inference to the Best Explanation. *Synthese*, *157*(3): 257–296.

Goldberg, S. C. (2012). Epistemic extendedness, testimony, and the epistemology of instrument-based belief. *Philosophical Explorations, 15*(2), 181-197. doi:10.1080/13869795.2012.670719

Goldberg, S.C. (2017). Epistemically engineered environments. *Synthese*,*197*, 2783–2802. doi:10.1007/s11229-017-1413-0

Goldenberg, M. J. (2021). *Vaccine Hesitancy: Public Trust, Expertise, and the War on Science* (H. E. Douglas Ed.): University of Pittsburgh Press.

Goldman, A. (1976). Discrimination and perceptual knowledge. *The Journal of Philosophy. 73*(20): 771-791. Doi:10.2307/2025679

Goldman, A. (1999). Internalism Exposed. *The Journal of Philosophy*, 96, 271–93.

Goldman, A. (2001). Experts: Which ones should you trust? *Philosophy and Phenomenological Research, 63*(1), 85-110. Retrieved from www.jstor.org/stable/3071090

Goldman, A. (2008). What is justified belief?, in Ernest Sosa, Jaegwon Kim, Jeremny Fantl and Matthew McGrath Eds. *Epistemology.* Malden: Blackwell.

Goldman, A. (2012). A Guide to Social Epistemology. In A. I. Goldman (Ed.). *Reliablism and Contemporary Epistemology: Essays*: Oxford University Press.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*: MIT Press.

Goodwin, J. (2011). Accounting for the Appeal to the Authority of Experts. *Argumentation, 25*, 285-296. doi:10.1007/s10503-011-9219-6

Graves, L., & Amazeen, M. A. (2019). Fact-Checking as Idea and Practice in Journalism. *Oxford Research Encyclopedia of Communications.* Retrieved from https://oxfordre.com/communication/view/10.1093/acrefore/9780190228613.001.0001/acrefore-9780190228613-e-808

Greco, J. (2012). Better safe than sensitive. In K. Becker & T. Black (Eds.), *The Sensitivity Principle in Epistemology* (pp. 193-206): Cambridge University Press.

Grimm, S. R. (2006). Is understanding a species of knowledge? *British Journal of Philosophy of Science, 57*, 515-535.

Grimm, S. R. (2010). The goal of explanation. *Studies in History and Philosophy of Science, 41*, 337-344. doi:10.1016/j.shpsa.2010.10.006

Grimm, S. R. (2016). Understanding and transparency. In S. R. Grimm, C. Baumberger, & S. Ammon (Eds.), *Explaining Understanding: New Perspectives from Epistemology and Philosophy of Science.* (pp. 212-229). New York: Routledge.

Grimsley, C. (2021). *Causal and Non-Causal Explanations of Artificial Intelligence*. Paper presented at the 27th Biennial Meeting of the Philosophy of Science Association, Baltimore, MD.

Guidotti, R., Anna, M., Ruggieri, S., Turini, F., Pedreschi, D., & Giannotti, F. (2018). A Survey of Methods for Explaining Black Box Models. arXiv:1802.01933v3

Gurumoorthy, K. S., Dhurandhar, A., & Cecchi, G. (2017). ProtoDash: Fast interpretable prototype selection. https://arxiv.org/abs/1707.01212

Haack, S. (1993). *Evidence and Inquiry: Towards Reconstruction in Epistemology*. Oxford: Blackwell Publishers.

Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. Available: arXiv:1903.03425.

Halpern, J. Y., & Pearl, J. (2005). Causes and explanations: A structural-model approach. Part I: Causes. *The British journal for the philosophy of science*, 56(4), 843-887.

Hampton, L. M. (2021). *Black Feminist Musings on Algorithmic Oppression*. Paper presented at the Conference on Fairness, Accountability, and Transparency (FAccT '21), Virtual Event. doi:10.1145/3442188.3445929

Hardwig, J. (1985). Epistemic Dependence. *The Journal of Philosophy, 82*(7), 335-349.

Hardwig, J. (1991). The role of trust in knowledge. *The Journal of Philosophy, 88*(12), 693-708. Retrieved from https://www.jstor.org/stable/2027007

Harman, G. (1973). *Thought*. Princeton: Princeton University Press.

Harrison, B., Ehsan, U., & Riedl, M. O. (2017). Rationalization: A Neural Machine Translation Approach to Generating Natural Language Explanations. *CoRR*. abs/1702.07826.

Hausman, D. M. (1992). Why look under the hood? In *Essays on philosophy and economic methodology* (pp. 70-73). Cambridge, UK: Cambridge University Press.

Hawley, K. (2017). Trustworthy Groups and Organizations. In P. Faulkner & T. Simpson (Eds.), *The Philosophy of Trust* (pp. 230-250). Oxford: Oxford University Press.

Hempel, C.G., & Oppenheim, P. (1948). Studies in the logic of explanation. *Philosophy of Science*, *15*(2), 135–175. Doi:10.1086/286983.

Hempel, C.G. (1965). *Aspects of scientific explanation and other essays in the philosophy of science*. New York: The Free Press.

Hinchman, E. (2005). Telling as Inviting to Trust. *Philosophy & Phenomenological Research, 70*, 562-587.

Holton, R. (1994). Deciding to trust, coming to believe. *Australasian Journal of Philosophy, 72*(1), 63–76.

Hubert, A., Bright, J., & Howard, P. N. (2020). *Social Media Junk News on Hydroxychloroquine and Trust in Science. University of Oxford.* Retrieved from https://comprop.oii.ox.ac.uk/wp-content/uploads/sites/93/2020/07/ComProp-Coronavirus-Misinformation-Weekly-Briefing-03-08-2020.pdf

Hume, D. (2007) [1748]. *An enquiry concerning human understanding. And other writings*. Cambridge: Cambridge University Press.

Humphreys, P. (2004). *Extending ourselves: Computational science, empiricism, and scientific method.* Oxford: Oxford University Press.

Humphreys, P. (2009). Network Epistemology. *Episteme, 6*(2), 221-229.

Iacovino, L. (2002). Ethical principles and information professionals: Theory, practice and education. *Australian Academic & Research Libraries, 33*(2), 57-74. doi:10.1080/00048623.2002.10755183

Ihde, D. (1990). *Technology and the Lifeworld: From Garden to Earth.* Bloomington: Indiana University Press.

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLoS Medicine*, *2(*8): e124. doi:10.1371/journal.pmed.0020124

Irving, G., & Askell, A. (2019). AI safety needs social scientists. *Distill*. doi:10.23915/distill.00014

Irving, G., Christiano, P., & Amodei, D. (2018). AI safety via debate. *arXiv stat.ML*. doi:arXiv:1805.00899

Irzik, G., & Kurtulmus, F. (2018). Well-ordered science and public trust in science. *Synthese*, *(198)*3: 4731–4748. doi:10.1007/s11229-018-02022-7

Irzik, G., & Kurtulmus, F. (2019). What is epistemic public trust in science. *British Journal of Philosophy of Science, 70*(4), 1145-1166.

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2013). *An introduction to Statistical Learning with applications in R*. First Edition: Springer.

Jebeile, J. and Kennedy, A. G. (2015). "Explaining with Models: The Role of Idealizations", *International Studies in the Philosophy of Science, 29*(4): 383–392. doi:10.1080/02698595.2015.1195143

Jiang, R., Chiappa, S., Lattimore, T., György, A., & Kohli, P. (2019). Degenerate feedback loops in recommender systems. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (pp. 383-390).

Jobin, A., Ienca, M. & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence,* 1, 389–399. doi:10.1038/s42256-019-0088-2

John, L. K., Loewenstein, G. & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23*(5), 524-532. doi:10.1177/0956797611430953

John, S. (2018). Epistemic trust and the ethics of science communication: against transparency, openness, sincerity and honesty. *Social Epistemology, 32*(2), 75-87. doi:10.1080/02691728.2017.1410864

John, S. (2021). Science, politics and regulation: The trust-based approach to the demarcation problem. *Studies in History and Philosophy of Science, 90*, 1-9. doi:10.1016/j.shpsa.2021.08.006

Jones, K. (1996). Trust as an Affective Attitude. *Ethics, 107*(1), 4–25. https://www.jstor.org/stable/2382241

Jones, K. (1998). Trust. In E. Craig (Ed.), *Routledge Encyclopedia of Philosophy* (pp. 466–470). London: Routledge. doi:10.4324/9780415249126-L107-1

Jones, K. (2004). Trust and Terror. In M. U. Walker & P. DesAutels (Eds.), *Moral psychology: Feminist ethics and social theory* (pp. 3-18). Lanham, MD: Rowman & Littlefield Publishers.

Jourová, V. (2020). Speech of Vice President Věra Jourová on countering disinformation amid COVID-19 "From pandemic to infodemic." Retrieved from https://ec.europa.eu/commission/presscorner/detail/en/speech_20_1000

Kant, I. (1993). On a supposed right to tell lies because of philanthropic concerns. In Kant, *Grounding for the Metaphysics of Morals*. Translated by J. Ellington, 3rd edn. Indianapolis, IN: Hackett.

Kant, I. (1996). *Metaphysics of Morals*. Translated by M. Gregor. In Kant, *Practical Philosophy*. Cambridge: Cambridge University Press.

Kant, I. (1997). *Lectures on Ethics*. Translated by P. Heath. Cambridge: Cambridge University Press.

Kahneman, D. (2011). *Thinking, fast and slow (1st ed).* New York: Farrar, Straus and Giroux.

Kato, M. P., Sakai, T. and Tanaka, K. (2013). When do people use query suggestions? A query suggestion log analysis. *Information Retrieval, 16*(6): 725–746.

Khalifa, K. (2013). Understanding, grasping, and luck. *Episteme, 10*(1), 1-17. doi:10.1017/epi.2013.6

Khalifa, K. (2017). *Understanding, Explanation, and Scientific Knowledge.* Cambridge: Cambridge University Press.

Kim, B., Rudin, C. & Shah, J. A. (2014). The Bayesian case model: A generative approach for case-based reasoning and prototype classification. in *Proc. Adv. Neural Inf.* 1952–1960.

Kitcher, P. (1990). The division of cognitive labor. *The Journal of Philosophy, 87*(1), 5-22. doi:131.111.5.152

Kitcher, P. (1993). *The Advancement of Science.* New York: Oxford University Press.

Klayman, J. (1995). Varieties of Confirmation Bias. *Psychology of Learning and Motivation, 32*, 385-418.

Köchling, A. & Wehner, M. C. (2020). Discrimination by an algorithm: a systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development. *Business Research, 13*(3), 795-848. doi:10.1007/s40685-020-00134-w

Kordig, C. R. (1978). Discovery and Justification. *Philosophy of Science, 45*(1), 110-117. doi:82.132.234.33

Kutrovatz, G., & Zemplen, G. A. (2011). Experts in Dialogue: An Introduction. *Argumentation, 25*(3), 275-283.

Krantz D. H., Peterson, N., Arora, P., Milch, K. & Orlove, B. (2008). Individual values and social goals in environmental decision making. In: Smith JC, Connolly T, Son YJ (eds.) Kugler T.

Decision modeling and behavior in complex and uncertain environments. New York, Springer, pp 165–198.

Krishnan, M. (2020). Against Interpretability: A Critical Examination of the Interpretability Problem in Machine Learning. *Philosophy & Technology, 33,* 487-502. doi:10.1007/s13347-019-00372-9.

Kuhn, M. & Johnson, K. (2013). *Applied Predictive Modeling.* New York, NY: Springer https://doi.org/10.1007/978-1-4614-6849-3.

Kuhn, T. (1962 [1970]). *The Structure of Scientific Revolutions*. Second Edition. Chicago: University of Chicago Press.

Kukla, R. (2012). "Author TBD": Radical Collaboration in Contemporary Biomedical Research. *Philosophy of Science, 79*(5), 845-858.

Kvanvig, J. (2003). *The value of knowledge and the pursuit of understanding*. Cambridge: Cambridge University Press.

Lackey, J. (1999). Testimonial knowledge and transmission. *The philosophical Quarterly, 49*(197), 471-490.

Lackey, J. (2006). It takes two to Tango Beyond reductionism and non-reductionism in the Epistemology of Testimony. In J. Lackey & Ernest Sosa (Eds.), *The Epistemology of Testimony*. (pp. 160-189). Oxford: Oxford University Press.

Lackey, J. (2008). *Learning from Words: Testimony as a Source of Knowledge*. Oxford: Oxford University Press.

Lackey, J. (2014). A Deflationary Account of Group Testimony. In Lackey (ed.). *Essays in Collective Epistemology*. Oxford: Oxford University Press.

Lackey, J. (2018). Group Assertion. *Erkenn, 83*(1), 21-42. doi:10.1007/s10670-016-9870-2

Lane, M. (2014). When the experts are uncertain: scientific knowledge and the ethics of democratic judgment. *Episteme, 2*(1), 97-118. doi:10.1017/epi.2013.48

Latour, B. & Woolgar, S. (1986). *Laboratory Life: The Construction of Scientific Facts*. second edition. Princeton, NJ: Princeton University Press.

Le Grand, J. (2003). *Motivation, Agency, and Public Policy: Of Knights and Knaves, Pawns and Queens*. Oxford: Oxford University Press.

Lehman, C.D., Yala, A., Schuster, T., Dontchos, B., Bahl, M., Swanson, K., & Barzilay, R. (2019). Mammographic Breast Density Assessment Using Deep Learning: Clinical Implementation. *Radiology, 290*, 52-58. https://doi.org/10.1148/radiol.2018180694

Lehrer, K. (2006). Testimony and trustworthiness. In J. Lackey & Ernest Sosa (Eds.), *The Epistemology of Testimony*. (pp. 145-159). Oxford: Oxford University Press.

Lehrer, K. & Wagner, C. (1981). *Rational Consensus in Science and Society.* Dordrecht: Reidel

Lemos, N. (2012). *An introduction to the theory of knowledge*. Cambridge: Cambridge University Press.

Leslie, D. (2019). *Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector*. Retrieved from The Alan Turing Institute.

Lewis, P., & Marsh, S. (2022). What is it like to trust a rock? A functionalist perspective on trust and trustworthiness in artificial intelligence. *Cognitive Systems Research, 72*, 33-49. doi:10.1016/j.cogsys.2021.11.001

Lipton, P. (1998). The Best Explanation of a Scientific Paper. *Philosophy of Science, 65*, 406–410.

Lipton, P. (2009). Understanding without explanation. In H. W. de Regt, S. Leonelli, & K. Eigner (Eds.), *Scientific understanding: Philosophical perspectives* (pp. 43–63). Pittsburgh: University of Pittsburgh Press.

Lipton, Z. C. (2017). The mythos of model interpretability. *2016 ICML Workshop on Human Interpretability in Machine Learning (WHI 2016)*. https://arxiv.org/abs/1606.03490

Locke, J. (1690). *An Essay Concerning Human Understanding*. London.

London, A. J. (2019). Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability. *The Hastings Center Report, 49*(1), 15-21. doi:10.1002/hast.973

Longino, H. E. (2002). *The Fate of Knowledge*. Princeton: Princeton University Press.

Ludwig, K. (2014). Proxy agency in collective action. *Nous*, *48*, 75–105.

Machamer, P., Darden, L. & Craver, C. F. (2000). Thinking about Mechanisms. *Philosophy of Science, 67*(1), 1-25. doi:10.1086/392759.

MacIntyre, A. (1995). Truthfulness, Lies, and Moral Philosophers: What Can We Learn from Mill and Kant?', in *The Tanner Lectures on Human Values*, Salt Lake City: University of Utah Press, 16, 307–361.

Mahon, J. (2006). Kant and the perfect duty to others not to lie. *British Journal for the History of Philosophy,* 14(4), 653–685.

Mahon, J. (2009). The Truth About Kant on Lies. In Clancy Martin (ed.), *The Philosophy of Deception*. (pp. 201-224). New York: Oxford.

Manson, N. C. (2010). Why do patients want information if not to take part in decision making? *Journal of Medical Ethics, 36*, 834-837. doi:10.1136/jme.2010.036491

Martini, C. (2014). Experts in science: a view from the trenches. *Synthese, 191*(1), 3-15.

Matheson, D. (2005). Conflicting experts and dialectical performance: Adjudication heuristics for the layperson. *Argumentation, 19*(2), 145-58.

Maynard, A. (2019). Ethics boards won't save big tech. Retrieved from: https://onezero.medium.com/tech-companies-need-an-ethics-reset-4d936a27960e

McMullin, E. (1968). "What Do Physical Models Tell Us?", in B. Van Rootselaar and J. Frits Staal (eds.), *Logic, Methodology and Philosophy of Science III* (Studies in Logic and the Foundations of Mathematics 52), Amsterdam: North Holland, pp. 385–396. doi:10.1016/S0049-237X(08)71206-0

McMullin, E. (1985). Galilean Idealization. *Studies in History and Philosophy of Science Part A*, 16(3): 247–273. doi:10.1016/0039-3681(85)90003-2

McMyler, B., & Ogungbure, A. (2018). Recent work on trust and testimony. *American Philosophical Quarterly*, *55*(3), 217–230. http://www.jstor.org/stable/45128616

Medawar, P. ([1963] 1996), "Is the Scientific Paper a Fraud?", reprinted in *The Strange Case of the Spotted Mice and Other Classic Essays on Science*. Oxford: Oxford University Press, 33–39.

Merton, R. K. (1973). *The Sociology of Science*. Chicago: University of Chicago Press.

Midden, C. J. H. & Huijts, N. M. A. (2009). The Role of Trust in the Affective Evaluation of Novel Risks: The Case of CO2 Storage. *Risk Analysis, 29*(5), 743–51.

Miller, B. (2013). When is consensus knowledge based? Distinguishing shared knowledge from mere agreement. *Synthese, 190*(7), 1293-1316. doi:10.1007/s11229-012-0225-5

Miller, B., & Record, I. (2013). Justified belief in a digital age: On the epistemic implication of secret internet technologies. *Episteme, 10*(2), 117-134. doi:10.1017/epi.2013.11

Miller, B., & Record, I. (2017). Responsible epistemic technologies: A social-epistemological analysis of autocompleted web search. *New Media & Society, 19*(12), 1945-1963. doi:10.1177/1461444816644805

Miller, T. (2018). Explanation in artificial intelligence: Insights from the social sciences. *arXiv:1706.07269v3*.

Misztal, B. A. (1996). *Trust in Modern Societies: Search for Bases of Social Order.* Cambridge, UK: Polity.

Mitchell, S. D., & Gronenborn, A. M. (2017). After fifty years, why are protein x-ray crystallographers still in business? *The British Journal of the Philosophy of Science. 68(3)*. 703-723.

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence, 1*, 501-507. doi:10.1038/s42256-019-0114-4

Moran, R. (2006). Getting told and being believed. In J. Lackey & Ernest Sosa (Eds.), *The Epistemology of Testimony*. (pp. 272 - 306). Oxford: Oxford University Press.

Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2020). From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices. *Science and engineering ethics*, *26*(4), 2141–2168. doi:10.1007/s11948-019-00165-5

Morrison, M. (2009). Models, measurement and computer simulation: The changing face of experimentation. *Philosophical Studies, 143*(1), 33-57.

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research (1979). *The Belmont report: Ethical principles and guidelines for the protection of human subjects of research.* U.S. Department of Health and Human Services. Retrieved from: https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/read-the-belmont-reprt/index.html

Nemitz, P. (2018). Constitutional democracy and technology in the age of artificial intelligence. *Philos. Trans. R. Soc.* A 376, 20180089.

Newman, L. (2019). Descartes' Epistemology. In E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy.* https://plato.stanford.edu/archives/spr2019/entries/descartes-epistemology/ (accessed March 3, 2022).

Nickel, P. J. (2011). Trust in technological systems. In M. J. de Vries, S. O. Hansson, & A. W. M. Meijers (Eds.), *Norms and the artificial: Moral and non-moral norms in technology*: Springer.

Nickel, P. J. (2012). Trust and testimony. *Pacific Philosophical Quarterly, 93*, 301-316. doi:10.1111/j.1468-0114.2012.01427.x

Nickel, P. J. (2013). Testimonial entitlement, norms of assertion and privacy. *Episteme, 10*(2), 207-217. doi:10.1017/epi.2013.17

Nickel, P. J. (2017). Being pragmatic about trust. In P. Faulkner & T. Simpson (Eds.), *The Philosophy of Trust* (pp. 195-213). Oxford Scholarship Online: Oxford University Press.

Nickel, P. J. (2021). Trust in engineering. In D. P. Michelfelder & N. Doorn (Eds.), *The Routledge Handbook of the Philosophy of Engineering* (pp. 494-505): Routledge.

Nickel, P. J., Franssen, M., & Kroes, P. (2010). Can we make sense of the notion of trustworthy technology? *Knowledge, Technology & Policy, 23*, 429-444.

Noble, S. U. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press.

Nozick, R. (1981). *Philosophical Explanations.* Cambridge, MA: Harvard University Press.

Nuremberg Code (1947). *Trials of war criminals before the Nuernberg Military Tribunals under Control Council law no.10 Nuernberg, October 1946 -April, 1949* . Washington, D.C., U.S.

Nyrup, R., & Robinson, D. (2022). Explanatory Pragmatism: context-sensitive framework for explainable medical AI. *Ethics and information technology*, 24(13). doi:10.1007/s10676-022-09632-3

O'Hara, K. (2020). Explainable AI and the philosophy and practice of explanation. *Computer Law & Security Review*, 39. doi:10.1016/j.clsr.2020.105474

O'Neill, O. (2002). *Autonomy and Trust in Bioethics.* Cambridge: Cambridge University Press.

O'Neill, O. (2007). A question of trust. (The BBC Reith lectures; 2002). Cambridge, UK: Cambridge University Press.

Ötting, S. K. & Maier, G. W. (2018). The importance of procedural justice in Human-Machine Interactions: Intelligent systems as new decision agents in organizations. *Computers in Human Behavior* 89: 27-39.

Páez, A. (2020). The pragmatic turn in explainable artificial intelligence (XAI). *Minds and Machines, 29*(3), 441-459. doi: 10.1007/s11023-019-09502-w.

Pariser, E. (2011). *The filter bubble: What the Internet is hiding from you*. London: Penguin.

Peirce, C. S. (1932). *Collected Papers of Charles Sanders Peirce*, vol. 2, C. Hartshorne and P. Weiss (eds.). Cambridge, MA: Harvard University Press.

Peters, D., Vold, K., Robinson, D., & Calvo, R. A. (2020). Responsible AI - Two frameworks for ethical design practice. *IEEE Transactions on Technology and Society, 1*(1), 34-47.

Pettit, P. (2002). Instituting a research ethic: Chilling and cautionary tales. In P. Pettit (Ed.), *Rules, Reasons, and Norms*. Oxford Scholarship Online: Oxford University Press.

Simon, J. (2013). Trust. In Pritchard, D. (ed.), *Oxford Bibliographies in Philosophy*, New York: Oxford University Press.

Simpson, T. (2012). What is trust? *Pacific Philosophical Quarterly, 93*, 550-569. doi:10.1111/j.1468-0114.2012.01438.x

Plantinga, A. (1993). *Warrant: The Current Debate*. Oxford: Oxford University Press.

Potochnik, A. (2016). Scientific explanation: Putting communication first. *Philosophy of Science, 83*(5), 721-732.

Price, W. N. II. (2015). Black Box Medicine. *Harvard Journal of Law & Technology, 28*(2), 419-467.

Pritchard, D. (2007). Anti-Luck Epistemology. *Synthese*, 158(3): 277–97.

Pritchard, D. (2005b). *Epistemic Luck*. Oxford: Oxford University Press.

Pritchard, D. (2008). Knowing the answer, understanding and epistemic value. *Grazer Philosophische Studien*, 77(1), 325-339. doi:10.1163/18756735-90000852

Putnam, H. (1981). *Reason, Truth and History.* Cambridge: Cambridge University Press.

Quine, W. and Ullian, J. (1970). *The Web of Belief*. New York: Random House.

Radder, H. (2003). Technology and Theory in Experimental Science. In H. Radder (ed.), *The Philosophy of Scientific Experimentation* (pp. 152-173). Pittsburgh: University of Pittsburgh Press, pp. 152–173.

Radder, H. (2006). *The World Observed/The World Conceived*. Pittsburgh, PA: University of Pittsburgh Press.

Reichenbach, H. (1938), *Experience and Prediction: An Analysis of the Foundations and the Structure of Knowledge.* Chicago: University of Chicago Press.

Reiss, J. (2008). *Error in economics: Towards a more evidence-based methodology*. New York: Routledge.

Reviving the US CDC. (2020). *The Lancet, 397*(10236): 1521. https://doi.org/10.1016/S0140-6736(20)31140-5

Ribeiro, M., Singh, S. & Guestrin, C. (2016). 'Why Should I Trust You?' Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM International Conference on Knowledge Discovery and Data Mining*. 1135-1144.

Rini, R. (2020). Deepfakes and the Epistemic Backstop. *Philosophers' Imprint, 20*(24), 1-16.

Rocke, A. (2010). *Image and reality: Kekulé, Kopp, and the scientific imagination*. Chicago, Ill.: University of Chicago Press.

Roetzel, P.G. (2019). Information overload in the information age: a review of the literature from business administration, business psychology, and related disciplines with a bibliometric approach and framework development. *Business Research, 12* (2), 479-522. https://doi.org/10.1007/s40685-018-0069-z

Rorty, R. (1979). *Philosophy and the Mirror of Nature*. Princeton, NJ: Princeton University Press.

Ross, C., & Swetlitz, I. (2017). IBM pitched its Watson supercomputer as a revolution in cancer care. It's nowhere close. *STAT*. Retrieved from: https://www.statnews.com/2017/09/05/watson-ibm-cancer/

Roush, S. (2005). *Tracking Truth: Knowledge, Evidence, and Science*. Oxford University Press.

Rudner, R. (1953). The Scientist *qua* Scientist Makes Value Judgments. *Philosophy of Science,* 20: 1–6.

Russell, S. and Norvig, P. (2010). *Artificial Intelligence: A Modern Approach* (3rd edn.). Upper Saddle River, NJ: Prentice Hall.

Salmon, W. (1984). *Scientific explanation and the causal structure of the world*. Princeton: Princeton University Press.

Schickore, J. (2008). Doing Science, Writing Science. *Philosophy of Science, 75*(3), 323-343. doi:131.111.098.148

Schmidt, S. (2009). Shall We Really Do It Again? The Powerful Concept of Replication Is Neglected in the Social Sciences. *Review of General Psychology, 13*(2), 90–100. doi:10.1037/a0015108

Scholz, O. R. (2009). Experts: What they are and how we recognize them - a discussion of Alvin Goldman's views. *Grazer Philosophische Studien, 79*(1), 187-205. doi:10.1163/18756735-90000864

Searle, J. R. (2001). *Rationality in Action*: The MIT Press.

Seger, E. (2022). In Defence of Principlism in AI Ethics and Governance. *Philosophy & Technology, 35*. doi:10.1007/s13347-022-00538-y

Seger, E., Avin, S., Pearson, G., Briers, M., Ó Heigeartaigh, S., & Bacon, H. (2020). *Tackling threats to informed decision-making in democratic societies: Promoting epistemic security in a technologically-advanced world*. Retrieved from https://www.turing.ac.uk/sites/default/files/2020-10/epistemic-security-report_final.pdf

Selbst, A. D., & Barocas, S. (2018). The intuitive appeal of explainable machines. *Fordham Law Review, 87*(3), 1085-1139. Retrieved from https://ir.lawnet.fordham.edu/flr/vol87/iss3/11

Shevlin, H. & Halina, M. (2019). Apply Rich Psychological Terms in AI with Care. Nature Machine Intelligence, 1(4): 165. https://www.nature.com/articles/s42256-019-0039-y?proof=t

Simmel, G. (1978 [1900]). *The Philosophy of Money.* London: Routledge.

Simmons, J., Nelson, L., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*(11), 1359–66. doi:10.1177/095679761141763

Simon, H. (1971). Designing Organisations for an Information Rich World. In M. Greenberger (Ed.) *Computer, communications, and public interest*. Baltimore, MD: The Johns Hopkins Press.

Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. [Online]. Available: https://arxiv.org/abs/1312.6034

Song, M., Yang, Y., He J., Yang, Z., Yu, S., Xie, Q. et al. (2018). Prognostication of chronic disorders of consciousness using brain functional networks and clinical characteristics. *eLife* 7:e36173. doi:10.7554/eLife.36173

Sosa, E. (1980), The Raft and the Pyramid: Coherence Versus Foundations in the Theory of Knowledge. *Midwest Studies in Philosophy*, *5*(1), 3–26.

Sosa, E. (1999a). How Must Knowledge Be Modally Related to What Is Known? *Philosophical Topics 26*(1/2): 373 –84.

Sosa, E. (1999b). How to Defeat Opposition to Moore. *Philosophical Perspectives, 13*: 141 –53.

Sosa, E. (2006). Knowledge: Instrumental and Testimonial. In J. Lackey & E. Sosa (Eds.), *The Epistemology of Testimony* (pp. 116–123). Oxford: Oxford University Press.

Stanley, D. (2003). What Do We Know about Social Cohesion: The Research Perspective of the Federal Government's Social Cohesion Research Network. *Canadian Journal of Sociology, 28*(1), 5–17.

Steup, M. (1996). An Introduction to Contemporary Epistemology. Upper Saddle River, NJ: Prentice-Hall.

Strevens, M. (2008). *Depth: An Account of Scientific Explanation*, Cambridge, MA, and London: Harvard University Press.

Strevens, M. (2016). How idealizations provide understanding. In S. R. Grimm, C. Baumberger, & S. Ammon (Eds.), *Explaining Understanding: New Perspectives from Epistemology and Philosophy of Science.* (pp. 35-49). New York: Routledge.

Sunstein C. R. (1996). On the expressive function of law. *University of Pennsylvania Law Rev*iew, *144*(5), 2021–2053.

Suppe, F. (1998). The Structure of a Scientific Paper. *Philosophy of Science 65*: 381–405.

Susskind, R., & Susskind, D. (2016). *The Future of the Professions: How Technology Will Transform the Work of Human Experts*. UK: Oxford University Press.

Taddeo, M. (2009). Defining trust and e-trust: From old theories to new problems. *International Journal of Technology and Human Interaction, 5*(2), 23-35.

Thagard, P. (1989). Explanatory coherence. *Behavioral and Brain Science. 12*(3), 435-467. doi: 10.1017/S0140525X00057046

Thagard, P. (2005). Testimony, credibility, and explanatory coherence. *Erkenntnis, 63*, 295-316. doi:10.1007/s10670-005-4004-2

Thompson, C. (2019). *Coders: Who they are, what they think, and how they are changing our world.* London: Pan Macmillan Press.

Tollefsen, D. P. (2007). Group Testimony. *Social Epistemology, 21*(3), 299-311. doi:10.1080/02691720701674163

Tomsett, R., Braines, D., Harborne, D., Preece, A., & Chakraborty, S. (2018). *Interpretable to whom? A role-based model for analyzing interpretable machine learning systems.* Paper presented at the 2018 ICML Workshop on Human Interpretability in Machine Learning (WHI 2018). https:// https://arxiv.org/abs/1806.07552

Trout, J. D. (2002). Scientific explanation and the sense of understanding. *Philosophy of Science, 69*(2), 212-233.

Trout, J. D. (2007). The psychology of scientific explanation. Philosophy Compass, pp. 564–591.

UK House of Lords Select Committee on Artificial Intelligence (2018). AI in the UK: ready, willing and able? HL Paper 100. https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/10002.htm

van Fraassen, B. C. (1980). *The Scientific Image*. Oxford: Clarendon Press.

van Wynsberghe, A., & Li, S. (2019). A paradigm shift for robot ethics: from HRI to human-robot-system interactions (HRSI). *Medicolegal and Bioethics, 9*, 11-21. doi:10.2147/MB.S160348

Vedder, A., & Naudts, L. (2017). Accountability for the use of algorithms in a big data environment. *International Review of Law, Computers & Technology, 31*(2), 206-224.

Verbeek, P. P. (2006). Materializing morality: Design ethics and technological mediation. *Science, Technology, & Human Values, 31*(3), 361-380. doi:10.1177/0162243905285847

Verbeek, P. P. (2008). Cyborg intentionality: Rethinking the phenomenology of human-technology relations. *Phenomenology and Cognitive Science, 7*(3), 387-395. doi:10.1007/s11097-008-9099-x

Vorvoreanu, M., & Walker, K. (2022). Advancing AI trustworthiness: Updates on responsible AI research. Retrieved from https://www.microsoft.com/en-us/research/blog/advancing-ai-trustworthiness-updates-on-responsible-ai-research/

Walton, N. D. (1989). Reasoned use of expertise in argumentation. *Argumentation, 3,* 59–73.

Walton, D. (1997). *Appeal to expert opinion: Arguments from authority.* University Park: Pennsylvania State University Press.

Ward. D., Hahn, J. and Feist, K. (2012). Autocomplete as a research tool: a study on providing search suggestions. *Information Technology and Libraries, 31*(4): 6–19.

Watcher, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanation without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law and Technology, 31*(2), 841-887.

Watson, D. (2019). The Rhetoric and Reality of Anthropomorphism in Artificial Intelligence. *Minds & Machines, 29*, 417-440. doi: 10.1007/s11023-019-09506-6

Weller, A. (2017). *Challenges for transparency*. Paper presented at the ICML Workshop on Human Interpretability in Machine Learning, Sydney, NSW, Australia.

Wheeler, B. (2020). Reliabilism and the testimony of robots. *Techne: Research in Philosophy of Technology, 24*(2).

White, W. W. and Marchionini, G. (2007). Examining the effectiveness of real-time query expansion. *Information Processing and Management, 43*(3): 685–704.

Whittlestone, J., Nyrup, R., Alexandrova, A., & Cave, S. (2019). The role and limits of principles in AI ethics: Towards a focus on tensions. *Proceedings of the 2019 AAI/ACM Conference on AI, Ethics, and Society.* https://doi.org/10.1145/3306618.3314289

Williamson, T. (2000). *Knowledge and its Limits.* Oxford: Oxford University Press.

Winsberg, E., Huebner, B., & Kukla, R. (2014). Accountability and values in radically collaborative research. *Studies in History and Philosophy of Science, 46*, 16-23. doi:http://dx.doi.org/10.1016/j.shpsa.2013.11.007

Woolley, A. W., Aggarwal, I., & Malone, T. W. (2015). Collective Intelligence and Group Performance. *Current Direction in Psychological Science, 24*(6), 420-424. doi:10.1177/0963721415599543

World Economic Forum, in collaboration with Deloitte and the Markkula Center for Applied Ethics at Santa Clara University (2020). *Ethics by Design: An organizational approach to responsible use of technology*. White Paper. Retrieved from http://www3.weforum.org/docs/WEF_Ethics_by_Design_2020.pdf (accessed March 17, 2021).

World Economic Forum, in collaboration with Deloitte and the Markkula Center for Applied Ethics at Santa Clara University (2021). *Responsible Use of Technology: The Microsoft Case Study*. White Paper. Retrieved from https://www3.weforum.org/docs/WEF_Responsible_Use_of_Technology_2021.pdf (accessed March 17, 2021).

World Health Organization. (2020). *Infodemic Management - Infodemiology*. Retrieved from https://www.who.int/teams/risk-communication/infodemic-management

Wortham, R., Theodorou, A., & Bryson, J. (2016). *What Does the Robot Think? Transparency as a Fundamental Design Requirement for Intelligent Systems*. Paper presented at the IJCAI 2016 Ethics for Artificial Intelligence Workshop, New York.

Wright, S. (2014). Sosa on Knowledge from Testimony. *Analysis, 74*(2), 249–254.

Zednik, C. (2019). Solving the black box problem: a normative framework for explainable artificial intelligence. *Philosophy & Technology,* 34, 265-288. doi:10.1007/s13347-019-00382-7

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. in *Proc. Eur. Conf. Comput. Vis*. Zurich, Switzerland: Springer. 818–833.

Zeng, Y., Lu, E. & Huangfu, C. (2019). Linking artificial intelligence principles. In *Proceedings of the 2019 AAAI Workshop on Artificial Intelligence & Safety (AAAI-Safe AI).* 1–15.

Zerilli, J., Knott, A., Maclaurin, J., & Gavaghn, C. (2018). Transparency in algorithmic and human decision-making: Is there a double standard? *Philosophy and Technology, 32*(4), 661-683. doi:https://doi.org/10.1007/s13347-018-0330-6

Zerilli, J. (2022). Explaining Machine Learning Decisions. *Philosophy of Science*, *89*(1), 1–19. https://doi.org/10.1017/psa.2021.13