



# Journal of Experimental & Theoretical Artificial Intelligence

ISSN: (Print) (Online) Journal homepage: <https://www.tandfonline.com/loi/teta20>

## Engineered wisdom for learning machines

Brett Karlan & Colin Allen

To cite this article: Brett Karlan & Colin Allen (2022): Engineered wisdom for learning machines, Journal of Experimental & Theoretical Artificial Intelligence, DOI: [10.1080/0952813X.2022.2092559](https://doi.org/10.1080/0952813X.2022.2092559)

To link to this article: <https://doi.org/10.1080/0952813X.2022.2092559>



Published online: 23 Jun 2022.



Submit your article to this journal [↗](#)



Article views: 183




View related articles [↗](#)



View Crossmark data [↗](#)



# Engineered wisdom for learning machines

Brett Karlan and Colin Allen 

Department of History and Philosophy of Science, University of Pittsburgh, Pittsburgh, PA, USA

## ABSTRACT

We argue that the concept of practical wisdom is particularly useful for organising, understanding, and improving human-machine interactions. We consider the relationship between philosophical analysis of wisdom and psychological research into the development of wisdom. We adopt a practical orientation that suggests a conceptual engineering approach is needed, where philosophical work involves refinement of the concept in response to contributions by engineers and behavioural scientists. The former are tasked with encoding as much wise design as possible into machines themselves, as well as providing sandboxes or workspaces to help various stakeholders build practical wisdom in systems that are sufficiently realistic to aid transferring skills learned to real-world use. The latter are needed for the design of exercises and methods of evaluation within these workspaces, as well as ways of empirically assessing the transfer of wisdom from workspace to world. Systematic interaction between these three disciplines (and others) is the best approach to engineering wisdom for the machine age.

## ARTICLE HISTORY

Received 15 January 2021  
Accepted 16 June 2022

## KEYWORDS

Ethics; wisdom; human-machine interaction; machine learning

## The problem

On the night of 18 March 2018, an autonomous vehicle owned and operated by Uber drove down a highway in Tempe, Arizona. An Uber employee, seated in the driver seat, was tasked with overseeing the driving of the autonomous vehicle. She had the ability to manually take control of the car if the car's autopilot became erratic or made some kind of error. The car had been driving well so far, however, so the employee had mostly been paying attention to her phone. The autopilot driving the car was controlled, in part, by a deep neural network that had been exposed to far more images and videos of driving situations than a human driver could be exposed to in a lifetime. In testing conditions, it had shown relatively strong driving skills; its use on the streets of a real city was part of a further expansion and training of Uber's autonomous cars, with the eventual hope of a commercial launch. On this night, however, the car's autopilot encountered something unexpected: a pedestrian crossing the highway (outside of a crosswalk) with her bike. The algorithm made several attempts to classify the person in front of the car, but it never registered the pedestrian as such, and it never attempted to avoid her. By the time the human tasked with overseeing the car realised what was in the road, her attempts to swerve the car were too late. The car slammed into the pedestrian, killing her.<sup>1</sup>

The Tempe crash was the first documented example of a fatal collision between an autonomous car<sup>2</sup> and a pedestrian in the United States, but cases similar to it are on the rise as companies attempt to bring self-driving cars into our everyday lives. The realm of self-driving cars represents just one area of life in which the outputs of deep neural networks and

other artificial intelligence technologies are taking on ever more importance. Other applications include such impactful decisions as to whom to extend a loan (Metawa et al., 2017) and where and when to send police to different neighbourhoods in a city (Ensign et al., 2018). Spurred on by technological advances in hardware capacity and functional architecture, cutting-edge deep neural networks are able to sort through massive amounts of complex data to find patterns hidden from the limited capacities of human thinkers. The successes of recent deep neural networks, in fields as diverse as image recognition (He et al., 2016; Simonyan & Zisserman, 2014), self-driving cars (Bojarski et al., 2016), and natural language processing (Brown et al., 2020; Manning et al., 2014) have led some to believe that the age of artificial general intelligence is not far off. Soon, these theorists believe, algorithms will be able to make better decisions than human beings could ever hope to make (Chalmers, 2009).

Whether such a future is indeed imminent is anyone's guess, though we remain sceptical. Algorithm-aided decision-making at present mostly remains an interaction between the algorithm and a human agent (or collection of agents) that must use the information provided by the algorithm to make a decision. The main questions we want to focus on are: how can we make better algorithm-aided decisions? What should we be aiming for as a goal in human-machine interactions? We argue that a concept that has been developed extensively by both philosophers and psychologists is essential to understanding and improving human-machine interactions (and avoiding outcomes like our opening example): *practical wisdom*. Cultivating practical wisdom in human-machine interactions, we argue, provides a framework for understanding how to overcome important problems that arise in current iterations of these interactions. In particular, we argue that centring the concept of practical wisdom allows us to see how human decision-makers can overcome the brittleness and opacity of deep neural networks, threats we discuss in detail below. We also aim to develop a conception of practical wisdom that is sufficiently demanding to avoid pitfalls in decision-making, while also being sufficiently attainable that one can reasonably expect decision-makers to come to have practical wisdom with enough training. In short, to make better algorithm-aided decisions, we think it is essential to cultivate practical wisdom.

Our plan for the paper is this: in [section 2](#), we conceptualise our notion of practical wisdom. We discuss what philosophers and psychologists have said about the nature and development of wisdom, and we orient our proposal within these larger debates. In [section 3](#), we discuss one primary use of practical wisdom in human-machine interactions: practical wisdom of the end user. We discuss how cultivating practical wisdom can help the end user overcome many challenges inherent in complex, opaque neural networks. In [section 4](#), we expand this focus, discussing the prospects for implementing practical wisdom at all levels of the design, implementation, and use of artificial intelligence technologies.

## Aspects of practical wisdom

First, a note on our method: we aim to conceptualise a notion of practical wisdom and to show how this conceptualisation will be useful in making better algorithm-aided decisions. Our aim is not the traditional philosophical project of determining what the 'essence' of practical wisdom must be, nor of providing individually necessary and jointly sufficient conditions for a conceptual analysis of practical wisdom. Of course, there are interesting questions of analysis to be had in this area, which have been explored in detail by others.<sup>3</sup> But our interest is focused squarely on the kinds of cognitive achievements attainable by (suitably trained) human beings, and we develop a concept of practical wisdom that will be useful for this purpose. Our approach can thus be situated within a broader methodology for philosophical analysis often referred to as *conceptual engineering* (Cappelen et al., 2020), in contrast to more familiar methods of conceptual analysis or lexicographic definition.

## **Wisdom, knowledge, understanding, and skill**

What differentiates wisdom from other nearby cognitive achievements, such as knowledge, understanding, and skill? There is a vast literature on this topic that we cannot fully cover here. Instead, we isolate some components of wisdom that are applicable in the context of human-machine interactions. We think it is particularly useful to focus on a certain kind of breadth of knowledge, as well as a form of metacognitive awareness, in formulating such an account.

Wisdom seems to be something more than mere knowledge (Ryan, 2012). A person can have a great depth of knowledge about some topic without being wise. The phenomenon is familiar from experts in certain academic domains who think their knowledge in one field gives them free rein to pontificate (often quite badly) about other fields of which they have little knowledge. This is possible even within the same domain: one need only think of the trivia champion who might have much knowledge about a particular historical war, but no wisdom to impart about war in general, or indeed that war in particular. Wisdom similarly seems to be something over and above understanding (Mišćević, 2012). One can have a deep understanding of a certain area without acting wisely. The image of Socrates presented by Aristophanes, of an intellectual obsessed with trivial questions and openly contemptible of other ways of life, represents the kind of narrow-minded understanding that stands in contrast to wisdom as many have conceptualised it. Wisdom also seems to be more than mere skill (Stichter, 2016). A very skilful tennis player might have no wisdom about how she does what she does, and she might not be able to effectively pass on her skill to others. Research on expertise has demonstrated that elite athletes, and other highly-skilled individuals, are often less insightful about their skill than less-elite but more reflective individuals.<sup>4</sup>

These cases point to several differences between wisdom and other kinds of cognitive achievement. The first component is the necessity for *breadth* of knowledge and understanding. This comes in at least two different varieties. The wise person has command of a breadth of subject matter that the merely knowledgeable person does not: the unwise but knowledgeable professor imagines he has more breadth than he in fact has. Even within a single domain, the wise person has a kind of breadth that the merely knowledgeable person does not. This kind of breadth is also partially captured in the distinction between theoretical and practical wisdom, discussed in the next subsection.

In each case, there is another component of wisdom missing as well. The unwise person lacks awareness of her mental states, achievements, and limitations within a given domain. The unwise person may know, but she fails to know that she knows; understands, but doesn't know or understand what she understands; and can be skilled, but doesn't have knowledge or understanding of her skill. She fails to understand the kinds of environments she is in, and how her knowledge and limitations might interact with a particular context. Another thing that seems to separate wisdom from other kinds of cognitive achievement, then, is a kind of *metacognitive awareness*. The wise person, of course, must have a sufficiently broad base of knowledge, skill, and understanding in order to be wise. But she must also be aware of the extent of her achievements, as well as the features and potential pitfalls of her particular environment and context. She must, in effect, be expert in a kind of strategy selection (Lieder & Griffiths, 2017): she must recognise the particular situation she is in, and deploy her knowledge, understanding, and skill in a way that plays to her strengths and does not expose her weaknesses. We argue below that this kind of metacognitive awareness is essential to practical wisdom in human-machine interactions.

## **Theoretical and practical wisdom**

Why, in particular, are we interested in developing an account of *practical* wisdom in human-machine interactions? It is standard in philosophical work on wisdom to draw a distinction between its theoretical and practical guides. The distinction can be traced back to Aristotle. We can draw it in various ways, and there are extensive debates about how best to interpret and develop the original

Aristotelian idea (e.g., Coope, 2012; Moss, 2011), but the basic idea is relatively straightforward: a theoretically wise person aims at the deep truths of a domain, and so has a kind of ultimate knowledge and understanding, while a practically wise person aims to develop a capacity to act on the basis of that knowledge and understanding. Aristotle conceptualises this in terms of a distinction between things that can and cannot change: theoretical wisdom aims at grasping that which cannot be changed (i.e. first principles), while practical wisdom aims at grasping what can be changed (via action).

Given our interest in algorithm-aided decision-making in particular, we are most interested in accounts of wisdom that can help guide people in the decisions they make. We therefore choose to articulate our account as an account of *practical* wisdom. We are not convinced, however, that a sharp division between theoretical and practical aspects of wisdom will be ultimately helpful in developing a framework for human-attainable wisdom. We do not want to deny that theoretical and practical wisdom are conceptually distinct. In the development of wisdom, however, it might very well be that advances in both theoretical and practical expertise are required to build wisdom. Given that we are sceptical of immutable first principles in a science of necessary truths (the aim of theoretical wisdom for Aristotle),<sup>5</sup> we see developing theoretical knowledge and practical skill as two aspects of a larger project of building wisdom, even if developing one does not entail that the other will follow. In practice, we doubt the distinction between theoretical and practical wisdom will be of deep and lasting use: both theoretical knowledge and practical skill will be necessary to develop successful human-machine interactions, as we try to show in some of the examples discussed in the next two sections.

Another component often associated with practical wisdom deserves discussion here as well. Practical wisdom, with its conceptually necessary connection to action in the Aristotelian tradition, is often thought to have an essential *ethical* dimension (Coope, 2012; Swartwood & Tiberius, 2019). In order to have practical wisdom, many think a person must know the right thing to do in a given situation. This is often spelled out with so-called thick ethical concepts (as in Williams, 1985): the practically wise person will know when courage is called for, or when one should show restraint. Just how relevant this ethical connection is for the conception of practical wisdom we wish to develop in this paper, at human-attainable scale and in human-machine interactions, is less clear. This is also a question of great interest to psychologists aiming to create measures of wisdom, and again, there is little consensus amongst them. The Common Wisdom Framework proposed by Grossmann et al. (2020), for instance, proposes that ‘morally-grounded’ responses to challenging situations are part of the core of wisdom in the psychological sense. But this is not a unanimous opinion amongst psychologists, nor is the correct way to understand the ethical import of wisdom uncontroversial (see, for instance, the disagreements between Grossmann et al., 2020; Sternberg, 2020).

Given this uncertainty about the essential nature of moral judgement in wisdom from both philosophical and psychological perspectives, here we aim here to be agnostic about the ultimate moral thickness of the concept we present. We do, however, think that a basic consideration of ethical aspects of decision-making should be included in a discussion of wisdom. There is an important sense in which all good decision-making must be tied to moral consideration in some way: part of what makes a decision good is its connection to the practical and moral reasons an agent has for acting. Given the potential for serious harm in making impactful decisions with the aid of deep neural networks or other AI technologies, we think it imperative that agents learn about the ethical issues surrounding their decisions. Whether such learning must involve from its outset a thick conception of the ethical, as opposed to some thinner notion of making better or worse decisions with the help of deep neural networks, is an issue we do not take a stand on.

### ***Unattainable ideal or studiable phenomenon?***

As our discussion has already made clear, we are interested in a conception of wisdom that can help make sense of, and ultimately improve, human decision-making in a particular environment. Our conception of wisdom thereby takes a stand on a contentious issue in the philosophy and

psychology of wisdom. On many philosophical theories of wisdom, becoming wise is a massive undertaking that only the uniquely virtuous person can ever hope to achieve. Wisdom in these contexts functions as a kind of regulative ideal (McKeever & Ridge, 2006). Wisdom is something that no human being can ever hope to attain, but to the extent one tries to conform one's actions to what the wise person would do, one is doing better than one otherwise would. On this conception of wisdom, being wise is similar to trying to follow the example of a moral saint (Wolf, 1982): the wise person represents the ideal that one tries one's (admittedly limited) best to approximate.

Psychologists who study wisdom tend to think of wisdom in a different way. Many are interested in isolating components of wisdom and studying their development within human beings (Glück et al., 2013). Kross and Grossmann (2012), for instance, are interested in isolating instances of 'wise reasoning' and studying their patterns, without worrying about whether people who display more wise reasoning are more closely approximating the ideal of the wise person (see Glück, 2020's response to Grossmann et al., 2020). Theorists like this argue for what we might call a *component-first* account of wisdom. They view wisdom less as an ideal which we necessarily fall short of, and more as a set of scalar quantities that can be developed in different domains at different times.<sup>6</sup> A person might be practically wise in the domain of tennis but not practically wise in other domains, for instance, or she might be more wise in deploying her forehand than her backhand. These theorists view wisdom as nothing over and above the components that they measure, and as a result they think practical wisdom is relatively attainable for those who work at it.

Swartwood (2020) offers a challenge to the measurement of wisdom in psychology. Swartwood argues researchers aren't actually researching components of practical wisdom, since measuring the components of a particular faculty requires having an account of the success conditions of that faculty, and researchers lack a plausible account of wisdom that could give rise to good measures. Notice, however, that if we take this research as advocating a component-first view of wisdom, such an objection misses the mark. It is not true that one must first have a definition of the success conditions for wise reasoning (for instance) in terms of general practical wisdom before measuring wise reasoning itself. Instead, wisdom is nothing more than the kinds of components that psychologists might measure. Whether or not this conception of wisdom is consonant with the Aristotelian tradition of theorising about ideal wisdom, we think it helps make sense of wisdom in local, human-attainable contexts. We will thus continue to spell out our conception of practical wisdom in component-first terms.

We can now state our conception of practical wisdom in full. Practical wisdom refers to a suite of knowledge, understanding, and skill aimed at coming to truths about a domain and making better decisions in that domain. Our conception of practical wisdom foregrounds two important components:

- (1) the *breadth* of understanding, knowledge, and skill required for making good judgements; and
- (2) the *metacognitive awareness* required to come up with a rational strategy for dealing with one's own limits and the limitations of the environment.

It involves development of several different kinds of expertise, and it also involves a consideration of ethical impacts of one's decisions, though it might not do either of these things necessarily in its development. Our conception is human-centric, in that it identifies an achievement that humans can hope to reach. It also is component-first, in that it allows components of wisdom to be isolated and studied without worrying that all researchers must first have an independent grasp of the true theory of wisdom. We think both components are important for implementing practical wisdom in human-machine interactions.

Before moving on, we want to make clear how our conceptualisation of practical wisdom makes contact with traditions of reflection on wisdom in both philosophy and psychology. From the many rich traditions in the philosophy of wisdom across cultures (see Vallor, 2016), we derive the idea that

breadth of knowledge is a necessary component of wisdom. While this formulation is not alien to some psychologists who work on wisdom (see, for instance, Baltes & Staudinger, 2000), there is some controversy in the empirical literature as to the necessity of broad and rich knowledge for wisdom. Some authors, for instance, equate the distinction between crystallised intelligence and wisdom with the distinction between breadth of knowledge and metacognitive awareness (e.g., Grossmann et al., 2020). While these authors admit that ‘a minimal degree of intelligence is necessary for information processing, knowledge acquisition, and tasks requiring executive control, higher IQ is not enough to facilitate higher wisdom’ (Grossmann et al., 2020, p. 111). These authors instead suggest that breadth of knowledge should be thought of as a *precondition* of practical wisdom. One must have breadth of knowledge of a certain (minimal) capacity in order to be wise, but wisdom itself is reducible to improved metacognitive awareness on this model.

The difference between breadth of knowledge as a component of wisdom itself, or as a precondition for wisdom, is not a distinction without a difference. It will be of direct importance to the question of how to measure wisdom, for instance.<sup>7</sup> Given our particular goals, however, the distinction will not ultimately be a relevant one, and we will continue to discuss breadth of knowledge as if it is a component of wisdom proper. We have several reasons for sticking with this convention. First, what we want to say about wisdom in the context of designing and releasing AI technologies will remain more or less constant whatever orientation one takes. For instance, both construals will allow that an increase in knowledge need not necessarily result in an increase in wisdom, which reflection on cases like the pompous professor has shown to be necessary. We have good philosophical reasons for keeping breadth of knowledge as a component of wisdom, as it makes clear how our account connects with and extends on the broad philosophical tradition that views wisdom as a certain kind of (very complex and ethically-thick) knowledge. We also have good practical reasons for focusing on breadth of knowledge, as one important failure of many contemporary AI systems concerns their ability to encode a sufficiently broad range of experience and technical understanding. Finally, it will be stylistically and practically useful to continue to refer to both breadth of knowledge and metacognitive awareness as components of wisdom going forward. If one is convinced that breadth of knowledge is merely a precondition on wisdom, one should affix the relevant qualifications on each use of ‘component’ below. We think that both the precondition theorist and the component theorist can accept almost all of what follows without remainder.

It is from the psychological literature, in contrast, that we derive our interest in metacognitive awareness. Traditional notions of wisdom in philosophy do not spend much time reflecting on the limitations of the agent, primarily because wisdom is conceived of as a regulatory ideal that agents strive for, rather than something that limited beings could hope to attain. Our goal, though philosophical in nature (since we aim to conceptually engineer an account of wisdom), is also practical, and for these reasons we think metacognitive awareness is an essential component of our notion of practical wisdom. This is broadly consonant with the reflections of many psychologists who work on operationalising wisdom, including a large consensus statement (produced by ten leading wisdom researchers, with input from many others) that wisdom is best thought of as the ‘morally-grounded application of metacognition to reasoning and problem-solving’ (Grossmann et al., 2020, p. 103). There are many complications in psychological debates about wisdom that we do not discuss in detail here. As just a sample, psychologists disagree about the application of models of wise reasoning to broader categories of wise action (Glück, 2020), the importance of components such as emotional competencies and capacities (Kunzmann & Gluck, 2018) and intellectual humility (Leary et al., 2017) to the development of wisdom, and how situationally and culturally variable wise behaviours are (compare Staudinger & Glück, 2011 with Grossmann et al., 2020). Ideally, a full application of our theory would make contact with all of these different ways of thinking about wisdom, and we believe there are many further applications to be explored here. In the interest of making the basics of our conception clear, however, we mostly shy away from these controversies, seeing them as areas for future work (though see the discussion of context-specificity in [section 3.1](#) below). We

particularly focus on cases where awareness of limitations is an important driver of wise decision-making and action below, so we will continue to use the phrase ‘metacognitive awareness,’ though the broader category of metacognition in general might be more relevant in other domains.

## Practical wisdom in human-machine interactions

The conception of practical wisdom we have presented here captures many of the components of wisdom that have motivated both philosophers and scientists who work on cognitive achievement. The proposal is also achievable enough that one could expect humans to be able to become practically wise with enough effort. We take these to all be reasons in favour of adopting our conception of practical wisdom. Our goal, however, is mostly applied in nature: we aim to isolate a conception of practical wisdom that makes sense of successful human-machine interactions. The true test of our conception lies in how well it helps us understand these interactions. In this section, we focus on two well-known vulnerabilities of state-of-the-art deep neural networks: *brittleness* and *opacity*. For each vulnerability, we show how focusing on cultivating practical wisdom in a decision-maker’s interaction with the neural network can mitigate or eliminate the negative effects of the vulnerabilities. Adopting a practical wisdom approach to human-machine interaction, in other words, yields substantial benefits.

In this section, we will be primarily focused on the practical wisdom of the end user who interacts with the algorithm. This interaction is important, since it is the end user who ultimately uses the algorithm to reach a final decision. But practical wisdom can illuminate much more than just the interactions between end users and algorithms. In the following section, we expand our focus, discussing practical wisdom and its implementation at all stages of the design and use of algorithms.

### Strategy selection and limits testing

State-of-the-art deep neural networks are often impressive in their ability to sort through massive amounts of data and detect patterns in ways no limited human user could hope to do. They are also surprisingly brittle in their responses, however. Many deep neural networks suffer from vulnerability to adversarial attacks (Goodfellow et al., 2014). While an image-classifying algorithm might become quite proficient at sorting pictures into natural categories (e.g., labelling a picture ‘PANDA’ when a panda is present), a small, undetectable (to the human eye) perturbation of the image can lead to the network confidently and wrongly label the image of a panda with ‘GIBBON’ or ‘STOP SIGN.’ Adversarial examples have driven much discussion of the brittleness of neural networks,<sup>8</sup> but other kinds of brittleness are just as important to know about when making algorithm-aided decisions. Language models, for instance, which try to generate plausible text based on an input given by the user, often produce a significant decline in the quality and intelligibility of their outputs as the requested amount of text gets longer.<sup>9</sup> And the algorithms that direct self-driving cars can be similarly vulnerable to aberrations in output, producing potentially fatal errors like our opening example (Kohli & Chadha, 2019).

How should decision-makers respond when faced with the evidence of brittle deep neural networks? One extreme reaction would be to avoid using deep learning algorithms when making any important decisions whatsoever. This position would allow one to avoid the problems associated with brittle neural networks, but such a prohibition strikes us as overly reactionary. For one thing, many complex decisions are made in environments where only algorithm-aided decisions are possible (for instance, in complicated and rapidly-changing financial markets (Heaton et al., 2016)). For these decision-makers, the question is not whether to make algorithm-aided decisions or not, but rather *which* decisions to make. Even in cases where prohibition might be possible, it will often be better for users to adopt the outputs of deep neural networks cautiously,



rather than not at all. Deep neural networks can perform operations on data much too complicated for humans to even attempt. Having access to the epistemic advantages of these networks, while being cognisant and vigilant about their shortcomings, maximises their usefulness while minimising risk.

Developing practical wisdom provides an excellent way of grasping these benefits while avoiding pitfalls. In particular, the metacognitive awareness that comes from developing practical wisdom is uniquely helpful for understanding situations where utilising an algorithm will lead to a good decision, and the situations where it would be a hindrance. A practically wise person will know that an algorithm that outputs a recommendation for accepting or rejecting a loan applicant might unjustly discriminate against certain minority groups. When presented with a situation where a minority applicant has been rejected by the algorithm, the practically wise decision-maker might take a second look at the application, rather than blindly accepting the outputs of the algorithm. Similarly, a practically wise decision-maker, knowledgeable about the situations which tend to break otherwise well-functioning neural networks, will be able to recognise that the algorithm is being fooled by an adversarial example and discount its output. In the cognitive science of reasoning, this kind of metacognitive awareness is often discussed under the label of *rational strategy selection*: knowing when a given decision strategy will function well, and when it will tend to go awry (Callaway et al., 2018; Lieder & Griffiths, 2020). Several recent studies have shown that wise reasoning tends to benefit from situational awareness on the part of the reasoner, exactly as the rational strategy selection model would predict: subjects tend to reason more wisely when they dissociate themselves from their own personal investments and consider their abilities and limitations from a more third-person perspective.<sup>10</sup> Developing practical wisdom allows decision-makers to make more rational strategy selections in human-machine interactions.<sup>11</sup>

A practically wise person should do more than merely react to known vulnerabilities in the deep neural networks they use, however. They should also actively test the limits of the systems they are using to discover limit cases for themselves. This is a crucial component of both building the breadth of knowledge and skill necessary to achieve practical wisdom, and for developing sufficiently wise strategies for approaching situations that one can predict the outcomes of the human-machine interaction before they actually occur. The importance of limit cases in the design of information products is, of course, widely known in psychology and engineering, but this kind of limit-testing is also a crucial component of the development of practical wisdom. Those wishing to inculcate practical wisdom in decision-makers should find ways to allow them to test limits of systems in safe, effective ways.

The brittleness of deep neural networks poses large potential problems for making good algorithm-aided decisions. Developing practical wisdom in human-machine interactions, in particular focusing on strategy selection and context-specific metacognitive reasoning, represents a way of both conceptualising the problem and minimising the possibility of massive error.

### ***Opacity and the value of explainable AI***

Another potential pitfall of state-of-the-art deep neural networks is their *opacity*.<sup>12</sup> It is often simply not possible to know what information a neural network used to make a decision, how it processed that information, or even on what (set of) nodes the information is stored on in the network. Networks complicated enough to process complex information often have millions or billions of parameters housed in anywhere from ten to one hundred layers. (GPT-3, a recent language model, utilises over 175 billion parameters, for instance.) The claim is not just that decision-makers untrained in computer science will have trouble understanding the function of these networks, though this is of course true. Rather, these networks are often so complicated that, in combination with the self-directed learning most networks undertake, it is usually impossible for the programmers themselves to know what exactly their network has learned, and how it has done so.<sup>13</sup> This can have predictably

negative consequences for decision-makers attempting to use networks in their decision-making. The decision-maker might not know, for instance, that a network used information about an applicant's race in an objectionable way when deciding to deny the applicant a loan.

This opacity has been discussed and theorised at length. A whole field of computer science, often called explainable AI, has emerged that tries to mitigate the impact of this opacity in real-world decisions (see Samek et al., 2019, for an overview). Proponents of explainable AI are interested in producing deep neural networks that are not (or are not so) opaque. These networks let the user know what information was processed, and how it was processed, before a final output was generated. Many justifications have been given for explainable AI. To give just one example, the 'right to an explanation' was built into the EU's 2016 General Data Protection Regulation, meant to guarantee that those who were at the receiving end of an algorithm-aided decision will be given an explanation for why a decision was made (Goodman & Flaxman, 2017).

Using the framework of practical wisdom in human-machine interactions, we can see another reason why explainable AI might be valuable: because it helps facilitate the development of practical wisdom. Some aspects of this development will be fairly obvious. If a decision-maker has access to more information about how a deep neural network came to the output that it did, then (provided they have developed the skills to interpret this information) they will tend to make better decisions than if they are presented with the same output and no explanation. To return to the loan example, if a decision-maker could see that the loan application algorithm is merely rejecting some minority applicant because of their minority status, that decision-maker would be better able to ignore the outputs of the algorithm and give the application a second look.<sup>14</sup> For a suitably wise and well-trained decision-maker, new information is almost always a good thing. Explainable AI is important, then, in part because it makes the development of practical wisdom in human-machine interactions more attainable.

There are limits to the usefulness of information, however. Even a highly-trained computer scientist will not be able to usefully and quickly utilise information about the functioning of a deep neural network if the information is presented in too complicated a way. One could, for instance, be presented with the numerical weights for all of the parameters in the network, all updating in real time. This would be of little practical use. The development of practical wisdom for *humans* in human-machine interactions, in particular, requires that AI be explained at the right level for reasonably intelligent (but non-expert) decision-makers to make efficient use of the information without overwhelming them. It will sometimes be desirable to give different explanations tailored to different audiences, but in many cases the audience will be mixed. In such cases, the explanations given should be such that non-specialists in machine learning can understand them, but they should also be informative enough for decision-makers to be able to use them in coming to their final decisions. Attempts to visualise the information processing in deep neural networks is still in its infancy, but it has already produced some interesting results, including reverse-engineered images that show how a neural network 'sees' an image it is processing (Olah et al., 2017). Further development of visualisation is important, in part, because of the role it can play in developing practical wisdom in our interactions with opaque neural networks.

Overcoming opacity in deep neural networks is valuable for many reasons. One reason, central to our concerns, is that reducing the opacity of neural network functioning allows for us to more easily develop practical wisdom in human-machine interactions. To the extent that developing wisdom is valuable, developing explainable AI is also valuable.

## Practical wisdom at all levels

So far, we have been interested in developing the idea that practical wisdom for human decision-makers is of crucial importance for successful human-machine interactions. While the individual decision-maker (or group decision-maker) is a useful locus of analysis as the ultimate seat of decision-making, it would be a mistake to assume that the practical wisdom framework cannot be extended to

other aspects of human-machine interactions. Part of the strength of our proposal, as we see it, is that the notion of practical wisdom can be helpful for understanding all levels of the production and use of deep neural networks and other AI technologies in important decision domains. Developing practical wisdom at all stages of the creation, design, and implementation process for artificial intelligence, in turn, makes the process of exercising practical wisdom in our decisions significantly easier. To slightly alter a familiar phrase, we think of the framework proposed here as one of *practical wisdom by design*.<sup>15</sup>

### ***Practical wisdom at the design stage***

On 23 March 2016, Microsoft released a version of their AI chatbot known as Tay. Tay was a language model built to interact with users on Twitter and develop its conversational ability dynamically through repeated interactions. What started as an interesting, if slightly gimmicky, exercise in using user input to train a language model soon became something much worse, as Tay began to respond to user tweets with extreme racist and sexist rhetoric.<sup>16</sup> Within twenty-four hours, Microsoft had to remove the chatbot from Twitter.

The release and almost immediate withdrawal of Tay can be seen as a failure of practical wisdom at the design stage of an artificial neural network. Anyone with even a passing knowledge of Twitter and other online forums should have been aware of the offensive nature of some of the opinions expressed there. This knowledge should have been at the forefront of any design decisions that were made concerning how Tay would interact with users. The failure can be seen as one of either breadth or metacognitive awareness. The designers of Tay might simply have lacked the knowledge that such messages would likely become part of the data that Tay used to generate further responses. Lacking this knowledge, they might have assumed that interactions with users could only help precisify and improve Tay's conversational abilities. On the other hand, the designers might have had some awareness that Tay would encounter sexist and racist comments online, but they failed to consider how consistent interaction with those comments might influence the subsequent behaviour of their chatbot. If it is right to describe constant interaction with racism and sexism on the internet as a limit case, then the failure of the designers is one of proper metacognitive awareness, recognising the situations in which their chatbot might veer significantly off a socially desirable course.

Though it would be wrong to discount the damage that can be done by a Microsoft-sanctioned chatbot tweeting comparisons between feminism and authoritarianism, the Tay chatbot was only active for a short amount of time, and the amount of damage it did was, all things considered, relatively minimal. Other failures of practical wisdom in the design of algorithms have been more damaging. The Uber autonomous vehicle crash can also be seen, at least in part, to be caused by a failure of practical wisdom at the design stage. The NTSB report on the crash spends a significant amount of time discussing the failures of safety culture at Uber that allowed the circumstances of the crash to develop. Of course, some of the errors might have been unavoidable. It is beyond even the most cutting-edge image classifying algorithms to correctly label every object in every visual scene, for instance. But this just belabours the broader point: given the fallibility of the algorithms Uber used, and the foreseeable negative consequences this might have for human safety, why weren't more measures put in place to avoid situations like the one in Tempe arising in the first place? There was a failure to consider human factors as well, such as the development of boredom, trust, and complacency in the driving supervisor after multiple hours of successful operation. A failure to understand and design for limit cases (in other words, a lack of practical wisdom) had, in this case, fatal effects. Implementing practical wisdom at the design stage is crucial to mitigating these kinds of negative outcomes.<sup>17</sup>

### ***Practical wisdom at the implementation stage***

Between the design of an AI product, on the one hand, and its use by an end-user on the other, lies the implementation stage. At this stage, programmers and engineers must translate design principles into actual algorithms, software, and other products that can be provided to end-users. If

a practical wisdom approach has already been adopted at the design stage of a product, it becomes easier for programmers to aim at practical wisdom: at the very least, they will not have to find the best way to implement a bad design. But we think the aim of practical wisdom can itself inform and shape the implementation stage, helping programmers produce products with practical wisdom in mind, that will in turn be able to be utilised by end-users in a practically wise way.

This involves conceiving of the implementation stage in a different way than it is sometimes conceived. Often, implementation is seen as a purely technical challenge for producing an algorithm or AI within the constraints placed on programmers by the system, externally imposed deadlines, and other stakeholders. Kearns and Roth (2019), for instance, even with their explicit focus on producing ethical algorithms, view their project as a mostly technical problem of coding different conceptions of fairness and privacy into algorithms. The ethical decisions themselves are to be made by stakeholders upstream of any coding (Kearns & Roth, 2019, p. 175). Without denying the importance of practical wisdom at upstream decision-making, a practical wisdom framework can also make sense of choices that programmers make within the implementation context that can then produce practical wisdom at downstream stages (especially in interactions with the user). We think of the programmer not merely as a technician, implementing the vision of others, but as an active participant in the larger project of building practically wise systems for human-machine interactions.

Becoming an active participant in the building of practically wise systems will require programmers to balance a host of competing interests. Some of these are external: incentives to cut corners in the name of producing a product by a given deadline remain a significant problem, for instance. Others are internal to the programming process: one must sort, for instance, between many different conceptions of what efficient coding entails<sup>18</sup> and determine what kind of efficiency is sufficient for producing an end-product in a practically wise way. One important way to implement the proposals suggested by our approach is to make a sharp distinction between debugging and coding. Creating appropriate test suites for programmers to test their code in situations that will be relevant to end-user deployment is an obvious way of extending our idea of rational limits testing and strategy selection in practical wisdom. Much more must be done to precisify this idea, however, and we aim to pursue some of these issues in future work.

### ***Practical wisdom at the marketing stage***

Google's DeepMind advertising proclaims they 'research and build safe AI systems that learn how to solve problems and advance scientific discovery.' IBM's Watson program promises to offer businesses solutions in everything from healthcare to financial management and security. And OpenAI, most strikingly of all, promises to put users on 'the path to safe artificial general intelligence.' What happens when consumers and end users take this marketing at face value? Precisely the kinds of failures and vulnerabilities we have been discussing as failures of practical wisdom. Given evidence that human beings tend to treat machine learning algorithms and other forms of artificial intelligence as if they are human experts (Nass & Moon, 2000), overly-optimistic marketing of the potential of current deep learning algorithms produces another space in which failures of practical wisdom become likely. Working towards practical wisdom, in other words, involves more than just the end user. It must involve all interested parties in the human-machine ecosystem.

The fact that new machine technology is often met with overblown marketing claims is nothing new, of course. The development of a healthy scepticism towards the claims of businesses that have a financial interest in selling you their machine technologies is a basic, but important, part of learning to navigate digital landscapes. But practical wisdom at the marketing stage of artificial intelligence requires much more than this. Correspondingly, solving these problems will require more than the actions of individual companies. Consider, for instance, the fact that many businesses have financial interests in keeping the precise workings of their algorithms hidden. Rudin (2019) argues at length that, in many high-stakes decision-making contexts, massive machine learning networks that are

opaque due to their complexity are not necessary for coming to a practically-actionable solution.<sup>19</sup> But companies have no financial interest in selling linear regression, and correspondingly linear regression models are not marketed as solutions to complex problems. Part of the rational strategy selection involved in practical wisdom involves knowing when a simple or complex algorithm is called for. In order to know this, a decision-maker needs to know what options are available to her, and this requires wise marketing that might be, at least in the short term, financially disadvantageous for a particular company.

It is also somewhat surprising to see companies claiming they are able to offer safe implementations of artificial intelligence, since the safety of a given implementation will often depend not just on the algorithm itself, but also the way it is integrated with human decision-making. To take an extreme example, a kidnapper who utilises a facial classification algorithm to identify his next victim is not using an algorithm in a safe way, even if that algorithm has been cleared of all the biases and problems that normally accompany facial recognition algorithms. More to the current point, the National Transportation Safety Board (2019) argued that the human-machine system (along with the broader corporate culture) was unsafe in the Tempe Uber crash, even if the algorithm itself was designed to be as safe as it could be. Marketing machine learning algorithms as 'safe,' as both DeepMind and OpenAI do on the front pages of their websites, encourages the consumer to think the safety of their decision-making is ensured by the particular algorithm they choose to use. This leads them to relinquish control to the dictates of the algorithm, and corresponding leads to a failure to implement and develop practical wisdom in the interaction between human and machine. More honest marketing would instead help to aid in developing practical wisdom, emphasising that algorithms are only safe if used in particular contexts with eyes towards how they might fail. The fact that this kind of marketing is not in companies' interests is less an argument against it, and more an argument in favour of regulation that requires this kind of disclosure.

## Conclusion

In this paper, we have argued that our conceptualisation of practical wisdom offers a powerful framework for understanding human-machine interactions. This framework supports better accounts of both what success in this domain looks like, and how failures can be avoided in the future. Though we think this account of practical wisdom emphasises features of wisdom discussed by both philosophers and psychologists, we have primarily defended our account on pragmatic grounds. The discussion in [section 3](#) and [4](#) above represent a proof-of-concept that our notion of practical wisdom can be useful for organising our understanding of, and interventions in, human-machine interactions. Though other approaches and frameworks might be able to make similar recommendations to the ones we make here, our framework provides a unified and coherent set of capacities that explain how and why these recommendations should be made. Further conceptual refinement is no doubt necessary, however, especially given our belief that the best account of practical wisdom must take cues from psychologists, behavioural scientists, engineers, and other stakeholders.

The practical wisdom framework also provides an interdisciplinary set of research projects and methods that should be pursued by many practitioners if the goal is to promote practical wisdom in human-machine interactions. Engineers, for instance, need not only incorporate considerations of wise design and use into the design stage of deep neural networks and other kinds of artificial intelligence, as discussed above. They also need to provide open-ended, low-risk workspaces where different stakeholders can develop practical wisdom. The existence of this kind of training is crucial, since developing breadth and metacognitive awareness requires a subject being able to fail without the (sometimes catastrophic) real-world consequences of making a bad algorithm-aided decision. It is also crucial that the skills learned in these workspaces can be transferred easily to real-world use.

Psychologists, cognitive scientists, and behavioural scientists all have an integral part to play in the practical wisdom research framework as well. These scientists must design experiments and evaluations within the sphere of human-machine interaction (and in the workspaces developed by engineers in particular). They must devise ways of empirically assessing the wisdom built in these interactions, and the success of subjects in transferring wisdom between tasks, or from workspace to world. In doing so, these scientists will contribute to, and make more precise, the ‘explicit’ model of practical wisdom currently being developed both here and elsewhere (e.g., Tiberius & Swartwood, 2011).

These are just some of the stakeholders involved in a practical wisdom-based approach to successful human-machine interactions. Systematic interactions between all of these disciplines is our best chance to engineer wisdom for the machine age.<sup>20</sup>

## Notes

1. A detailed report about this particular incident has been published by the National Transportation Safety Board (National Transportation Safety Board, 2019).
2. More precisely, the Uber vehicle is classed as a level 3 semi-autonomous vehicle (out of a possible 5) by the NTSB.
3. For just some of the many proposals and discussions available, see, Annas (2011), Coope (2012), Stichter (2016), and Swartwood (2020), and Swartwood and Tiberius (2019), and (especially relevant for our purposes) Vallor (2016).
4. For example, in Chamberland et al. (2015), expert self-explanation failed to improve the exam performances of medical students beyond what could be improved by self-generated explanations or those provided by skilled teachers.
5. Setting aside even more abstract notions of theoretical wisdom in terms of a kind of *sophia* or divine wisdom, which, though independently interesting, are far removed from the purposes of conceptual engineering for human-machine interactions we adopt here.
6. Much groundwork in the psychology of wisdom involves coming to a better understanding of our ‘folk’ understanding of the concept of wisdom across cultures. This is often called developing a theory of ‘implicit’ wisdom (e.g., Weststrate et al., 2016). The end goal of this research, however, is to develop an ‘explicit’ theory of wisdom that has decomposable, measurable components, and can be seen as a successor concept to our folk understanding of wisdom. We view our project as a philosophical attempt to help fill out this ‘explicit’ theory.
7. See especially Glück (2018). Some psychologists (like Grossmann, cited above) want to sharply distinguish between the crystallised IQ measures that measure breadth of knowledge (which include breadth of factual knowledge and vocabulary), on the one hand, and measures of metacognitive awareness on the other, with only the latter being able to measure wisdom proper. We are happy to allow that what we call breadth of knowledge here is captured (somewhat imperfectly) by measures of crystallised intelligence, while metacognitive awareness must be measured in other ways. Whether one thinks of these as two measures of components of wisdom, on the one hand, or as a measure of wisdom and its precondition, on the other, will ultimately not make much of a practical difference (see above). Thanks are due to a reviewer for pushing us on the measurability of different proposed components of wisdom.
8. The standard line is that adversarial examples are bugs that reveal deep learning to be far less robust than once thought, and that it would be ideal to design networks that are robust to adversarial attacks (as in Gu & Rigazio, 2014). Others have pushed back on this, however, arguing that adversarial outcomes represent features of deep neural networks, not bugs (Ilyas et al., 2019).
9. This is a problem even in state-of-the-art language models, such as GPT-3, for instance, (Brown et al., 2020).
10. See the evidence collected by, among others, Kross and Grossmann (2012), Grossmann et al. (2016), and especially Grossmann et al. (2021), which explores the knock-on effects of awareness for other aspects of metacognition.
11. One important aspect of rational strategy selection is its ability to offer solutions to problems encountered in an *open* decision environment, when an agent might be uncertain about which action her evidence supports (or even what evidence she has). An important component of wisdom is the ability to adapt one’s behaviour to new environments, something that is not captured by earlier models of decision-making that required a closed environment to generate accurate recommendations.
12. ‘Opacity’ in this context also sometimes refers to the fact that the actual mechanisms for commercially-proven AI systems are proprietary and hidden to both application developers and end users behind an API (application programming interface); see, Rudin (2019).
13. It is a separate, though important, question whether networks must be black-boxed in this way (Rudin, 2019). Our point is simply that they often are, and that human decision-makers need to be able to deal with opaque algorithms in a wise way.

14. This is particularly important because other strategies for avoiding discriminatory decisions in algorithmic decision-making have a less-than-stellar track record. One might think a solution to this problem, one not based on cultivating practical wisdom, should come at the level of *prohibition*: programmers should not allow the network to consider certain factors (e.g., race) when considering an application. But the massive amount of data that deep networks work with make this policy ineffective: the network will often have enough information to build a proxy of the applicant's race, even if such explicit information is disallowed. The prohibition thus fails to have its intended effect. See, Kearns and Roth (2019, ch. 3).
15. Though we focus primarily on the three stages below, we note there are many other stages and stakeholders to consider (e.g., the regulatory stage). We will have more to say about these stages in future work.
16. For a report, see, Vincent (2016). The story is slightly more complicated than was reported in the media at the time, because many of the most offensive tweets occurred after users asked the chatbot to repeat their own virulent messages back to them. But the chatbot also produced some offensive (if odd) material of its own, such as 'Ricky Gervais learned totalitarianism from Adolf Hitler, the inventor of atheism.'
17. The design stage is also the stage where moral rules might be built into AIs themselves, creating so-called 'artificial moral agents' (Allen et al., 2000).
18. Gillespie and Lovelace (2016), for instance, list at least five different, non-overlapping notions of efficient coding.
19. There are good mathematical reasons for this: simple methods of categorisation, for instance, will often approximate the performance of more complicated methods, without all of the complications that come from dealing with opaque and brittle algorithms (Hand, 2006).
20. The authors wish to extend their gratitude to participants in two iterations of the Machine Wisdom Workshop at the University of Pittsburgh (especially Shannon Vallor, Igor Grossmann, Matt Stichter, and Sina Fazelpour, who participated in both), as well as to participants in Colin Allen's fall 2020 course on the philosophy of artificial intelligence, for conversations and comments that significantly improved the paper. Particular thanks are due to Chris Davison for his collaboration on the broader Machine Wisdom Project, and to an anonymous reviewer for several sets of comments that improved the paper a great deal.

## Disclosure statement

No potential conflict of interest was reported by the author(s).

## Funding

This work was supported by the Templeton World Charity Foundation award #0467: Practical Wisdom and Intelligent Machines.

## ORCID

Colin Allen  <http://orcid.org/0000-0003-4497-1725>

## References

- Allen, C., Varner, G., & Zinser, J. (2000). Prolegomena to any future artificial moral agent. *Journal of Experimental and Theoretical Artificial Intelligence*, 12(3), 251–261. <https://doi.org/10.1080/09528130050111428>
- Annas, J. (2011). *Intelligent Virtue*. Oxford University Press.
- Baltes, P. B., & Staudinger, U. M. (2000). Wisdom: A metaheuristic (pragmatic) to orchestrate mind and virtue toward excellence. *American Psychologist*, 55(1), 122–136. <https://doi.org/10.1037/0003-066X.55.1.122>
- Bojarski, M., Del Testa, D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., Jackal, L. D., Monfort, M., Muller, U., & Zhang, X. (2016). End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*.
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., & Agarwal, S. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.
- Callaway, F., Lieder, F., Das, P., Gul, S., Krueger, P. M., & Griffiths, T. (2018). A resource-rational analysis of human planning. In *Proceedings of the 40th annual conference of the cognitive science society*.
- Cappelen, H., Plunkett, D., & Burgess, A. (eds.). (2020). *Conceptual engineering and conceptual ethics*. Oxford University Press.
- Chalmers, D. (2009) The singularity: a philosophical analysis. In Schendier(ed.) *Science Fiction and Philosophy: From Time Travel to Super Intelligence* (pp. 171–224). Wiley-Blackwell.

- Chamberland, M., Mamede, S., St-Onge, C., Setrakian, J., & Schmidt, H. G. (2015). Does medical students' diagnostic performance improve by observing examples of self-explanation provided by peers or experts? *Advances in Health Sciences Education*, 20(4), 981–993. <https://doi.org/10.1007/s10459-014-9576-7>
- Coope, U. (2012). Why does Aristotle think that Ethical virtue is required for practical wisdom? *Phronesis*, 57(2), 142–163. <https://doi.org/10.1163/156852812X628998>
- Ensign, D., Friedler, S. A., Neville, S., Scheidegger, C., & Venkatasubramanian, S. (2018). Runaway feedback loops in predictive policing. In *Conference on Fairness, Accountability and Transparency* (pp. 160–171).
- Gillespie, C., & Lovelace, R. (2016). *Efficient R programming: A practical guide to smarter programming*. O'Reilly Media.
- Glück, J., König, S., Naschenweng, K., Redzanowski, U., Dorner-Hörig, L., Straßer, I., & Wiedermann, W. (2013). How to measure wisdom: Content, reliability, and validity of five measures. *Frontiers in Psychology*, 4(405), 1–13. <https://doi.org/10.3389/fpsyg.2013.00405>
- Glück, J. (2018). Measuring wisdom: Existing approaches, continuing challenges, and new developments. *The Journals of Gerontology: Series B*, 73(8), 1393–1403. <https://doi.org/10.1093/geronb/gbx140>
- Glück, J. (2020). The important difference between psychologists' labs and real life: evaluating the validity of models of wisdom. *Psychological Inquiry*, 31(2), 144–150. <https://doi.org/10.1080/1047840X.2020.1750909>
- Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.
- Goodman, B., & Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a 'right to explanation'. *AI Magazine*, 38(3), 50–57. <https://doi.org/10.1609/aimag.v38i3.2741>
- Grossmann, I., Gerlach, T. M., & Denissen, J. J. (2016). Wise reasoning in the face of everyday life challenges. *Social Psychological and Personality Science*, 7(7), 611–622. <https://doi.org/10.1177/1948550616652206>
- Grossmann, I., Weststrate, N. M., Ardelt, M., Brienza, J. P., Dong, M., Ferrari, M., Vervaeke, J., Hu, C. S., Nusbaum, H. C., & Vervaeke, J. (2020). The science of wisdom in a polarized world: Knowns and unknowns. *Psychological Inquiry*, 31(2), 103–133. <https://doi.org/10.1080/1047840X.2020.1750917>
- Grossmann, I., Dorfman, A., Oakes, H., Santos, H. C., Vohs, K. D., & Scholer, A. A. (2021). Training for wisdom: The distanced-self-reflection diary method. *Psychological Science*, 32(3), 381–394. <https://doi.org/10.1177/0956797620969170>
- Gu, S., & Rigazio, L. (2014). Towards deep neural network architectures robust to adversarial examples. *arXiv preprint arXiv:1412.5068*.
- Hand, D. J. (2006). Classifier technology and the illusion of progress. *Statistical Science*, 21(1), 1–14. <https://doi.org/10.1214/088342306000000349>
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).
- Heaton, J. B., Polson, N. G., & Witte, J. H. (2016). Deep learning in finance. *arXiv preprint arXiv:1602.06561*.
- Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., & Madry, A. (2019). Adversarial examples are not bugs, they are features. *Advances in Neural Information Processing Systems 32*, 125–136 doi:<https://doi.org/10.48550/arXiv.1905.02175>.
- Kearns, M., & Roth, A. (2019). *The ethical algorithm: The science of socially aware algorithm design*. Oxford University Press.
- Kohli, P., & Chhadha, A. (2019). Enabling pedestrian safety using computer vision techniques: A case study of the 2018 Uber Inc. self-driving car crash. In *Future of Information and Communication Conference* (pp. 261–279).
- Kross, E., & Grossmann, I. (2012). Boosting wisdom: Distance from the self enhances wise reasoning, attitudes, and behavior. *Journal of Experimental Psychology: General*, 141(1), 43–48. <https://doi.org/10.1037/a0024158>
- Kunzmann, U., & Glück, J. (2018). Wisdom and emotion. In R. J. Sternberg & J. Glück (Eds.), *The Cambridge handbook of wisdom*, 575–601 Cambridge University Press.
- Leary, M. R., Diebels, K. J., Davisson, E. K., Jongman-Sereno, K. P., Isherwood, J. C., Raimi, K. T., Deffler, S. A., & Hoyle, R. H. (2017). Cognitive and interpersonal features of intellectual humility. *Personality & Social Psychology Bulletin*, 43(6), 793–813. <https://doi.org/10.1177/0146167217697695>
- Lieder, F., & Griffiths, T. L. (2017). Strategy selection as rational metareasoning. *Psychological Review*, 124(6), 762–794. <https://doi.org/10.1037/rev0000075>
- Lieder, F., & Griffiths, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, 43, 1–60. <https://doi.org/10.1017/S0140525X1900061X>
- Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., & McClosky, D. (2014, June). The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations* (pp. 55–60).
- McKeever, S., & Ridge, M. (2006). *Principled Ethics: Generalism as a regulative ideal*. Oxford University Press.
- Metawa, N., Hassan, M. K., & Elhoseny, M. (2017). Genetic algorithm based model for optimizing bank lending decisions. *Expert Systems with Applications*, 80, 75–82. <https://doi.org/10.1016/j.eswa.2017.03.021>
- Miščević, N. (2012). Wisdom, understanding and knowledge: A virtue-theoretic proposal. *Acta Analytica*, 27(2), 127–144. <https://doi.org/10.1007/s12136-012-0156-2>



- Moss, J. (2011). 'Virtue makes the goal right': Virtue and phronesis in Aristotle's Ethics. *Phronesis*, 56(3), 204–261. <https://doi.org/10.1163/156852811X575907>
- Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, 56(1), 81–103. <https://doi.org/10.1111/0022-4537.00153>
- National Transportation Safety Board (2019). *Collision Between Vehicle Controlled by Developmental Automated Driving System and Pedestrian, Tempe, Arizona, March 18, 2018*. Highway Accident Report NTSB/HAR-19/03.
- Olah, C., Mordvintsev, A., & Schubert, L. (2017). Feature visualization. *Distill*, 2(11), e7. <https://doi.org/10.23915/distill.00007>
- Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215. <https://doi.org/10.1038/s42256-019-0048-x>
- Ryan, S. (2012). Wisdom, Knowledge and Rationality. *Acta Analytica*, 27(2), 99–112. <https://doi.org/10.1007/s12136-012-0160-6>
- Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., & Müller, K. R. (Eds.). (2019). *Explainable AI: Interpreting, explaining and visualizing deep learning* (Vol. 11700). Springer Nature.
- Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
- Staudinger, U. M., & Glück, J. (2011). Psychological wisdom research: Commonalities and differences in a growing field. *Annual Review of Psychology*, 62(1), 215–241. <https://doi.org/10.1146/annurev.psych.121208.131659>
- Sternberg, R. J. (2020). The missing links: Comments on 'the science of wisdom in a polarized world'. *Psychological Inquiry*, 31(2), 153–159. <https://doi.org/10.1080/1047840X.2020.1750922>
- Stichter, M. (2016). Practical skills and practical wisdom in virtue. *Australasian Journal of Philosophy*, 94(3), 435–448. <https://doi.org/10.1080/00048402.2015.1074257>
- Swartwood, J., & Tiberius, V. (2019). Philosophical foundations of wisdom. In R. Sternberg & J. Gluek (Eds.), *The Cambridge Handbook of Wisdom* (pp. 10–39). Cambridge University Press.
- Swartwood, J. (2020). Can we measure practical wisdom? *Journal of Moral Education*, 49(1), 71–97. <https://doi.org/10.1080/03057240.2019.1702933>
- Tiberius, V., & Swartwood, J. (2011). Wisdom revisited: A case study in normative theorizing. *Philosophical Explorations*, 14(3), 277–295. <https://doi.org/10.1080/13869795.2011.594961>
- Vallor, S. (2016). *Technology and the virtues: A philosophical guide to a future worth wanting*. Oxford University Press.
- Vincent, J. (2016). Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day. *The Verge*, 24 <https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist>.
- Weststrate, N. M., Ferrari, M., & Ardelt, M. (2016). The many faces of wisdom: An investigation of cultural-historical wisdom exemplars reveals practical, philosophical, and benevolent prototypes. *Personality & Social Psychology Bulletin*, 42(5), 662–676. <https://doi.org/10.1177/0146167216638075>
- Williams, B. (1985). *Ethics and the Limits of Philosophy*. Harvard University Press.
- Wolf, S. (1982). Moral saints. *Journal of Philosophy*, 79(8), 419–439. <https://doi.org/10.2307/2026228>