Edited Book Title: ‘Overcoming the Myth of Neutrality: Expertise for a New World’

Chapter Title:

Trustworthy experts and untrustworthy experts: Insights from the cognitive psychology of expertise

Andrew J. Waters

Fernand Gobet

Abstract

A central concern of the book is the importance to being able to distinguish “trustworthy subjects – who increase the epistemic welfare of our communities – from those untrustworthy individuals, who instead do not deserve our epistemic credit”. We address this question from the perspective of theory and data from the cognitive psychology of expertise. We will argue that trustworthy subjects, or “true experts”, are different from “untrustworthy subjects” in three ways. First, true experts can only develop in a regular environment in which there are reliable relationships between cues and outcomes that are “learnable”, whether by humans or algorithms. We will note that some environments are much more regular than others, and therefore provide a foundation for expertise to potentially develop. Note that this feature is a property of the environment, rather than the individual. Second, individuals need to have the opportunity to learn relationships between cues and outcomes. This will usually require the presence of a kind learning environment. A kind learning environment involves the receipt of continuous, fast, and accurate feedback from the environment. Third, individuals will need the opportunity to engage in long periods of structured practice. Over time, performance becomes more accurate (less bias) and less variable (less noise). Of course, there may be individual differences in learning rates in the same learning environment. In some domains, there is little ambiguity as to what constitutes accuracy, but in other domains the context can make accuracy harder to define. We also compare human experts with machine experts in a number of domains. Overall, true experts can be found in regular learning environments when they have had the opportunity to learn relationships over a long period of time, and these individuals, who are accurate and consistent, deserve our epistemic credit.

Introduction

One central concern of the edited volume is the importance to being able to distinguish "trustworthy subjects – who increase the epistemic welfare of our communities – from those untrustworthy individuals, who instead do not deserve our epistemic credit". A second concern is the question of whether possessing what we call "true expertise" (or "epistemic authority by virtue of reference to scientific criteria and methods") can automatically make an individual's assertions necessarily more grounded than others. A third concern, as revealed in the book's title, is how expertise is viewed in the post-pandemic world, given competing inputs from human experts and other stakeholders who may weigh variables differently, as well as the role of "non-human experts" in the form of computer algorithms.

In this chapter, we address these three questions from the perspective of theory and data from the cognitive psychology of expertise. We will argue that trustworthy subjects, or "true experts", are different from "untrustworthy subjects", in three ways. First, true expertise can only develop in a regular environment in which there are reliable relationships between cues and outcomes that are "learnable", whether by humans or algorithms. We will note that some environments are much more regular than others, and therefore provide a foundation for expertise to potentially develop. Note that this feature is a property of the environment, rather than the learner. Second, individuals need to have the opportunity to learn relationships between cues and outcomes. This will require a kind learning environment. A kind learning environment involves the receipt of continuous, fast and accurate feedback from the environment. Third, future true experts need to have the motivation to engage in sufficient structured practice, such as "deliberate practice" (Ericsson et al., 1993).

This chapter draws on ideas from the cognitive psychology literature on expertise to include those expressed in Shanteau (1992a, b), Kahneman (2011), Gobet (2016), and Kahneman et al. (2021). Therefore, the material and perspective presented in this chapter is not

novel. However, the contribution of this chapter is the application of these ideas to the framework of the book.

Note on Terminology

In this chapter, we shall use terminology from cognitive psychology. So, we will start by defining some terms. We refer to expertise as the learning of relationships between cues and outcomes. In this context, we use the term “cue” in a broad sense to refer to patterns of stimuli in the environment. We use the term “outcome” to refer to decisions/actions taken in response to cues. We use the term “feedback” to refer to the visible result of the decisions/actions that are taken. This can involve objective feedback from reality (winning or losing a chess match) or subjective feedback from other humans. Regarding the cognitive processes underlying decisions/actions, we use the term “chunk” to refer to the mental representation of cues in the environment. When we refer to performance with “low bias”, we mean performance that is generally accurate, and when we refer to performance with “low noise”, we mean performance which does not vary much from trial to trial, that is, performance which is consistent.

Finally, as described by Gobet (2016), there is no consensus regarding the definition of “expertise”. In this chapter we will use the definition provided by Gobet (2016, p. 5) and define an expert as “somebody who obtains results that are vastly superior to those obtained by the majority of the population”. As noted by Gobet (2016), this definition has the advantage that it can be applied to different subpopulations, and that we can define a “super-expert”, such as a chess super-grandmaster, as somebody whose performance is generally superior to the majority of grandmasters.

1. Trustworthy vs Untrustworthy Subjects

From the perspective of cognitive psychology, there are three dimensions to be considered. The first dimension concerns the properties of the task environment. We will note that some environments are much more regular than others, and therefore provide a foundation for expertise to potentially develop. Note that this feature is a property of the environment, rather

than the individual. True expertise can only develop in a regular environment in which there are reliable relationships between cues and outcomes that are “learnable”, whether by humans or algorithms.

The second dimension is the learning environment. Specifically, individuals need to have the opportunity to learn relationships between cues and outcomes. Different types of learning environments can either promote or inhibit learning of the relationships between cues and outcomes. Learning will be enhanced by a kind learning environment, which involves the receipt of continuous, fast, and accurate feedback from the environment. Learning will be inhibited in an unkind learning environment in which feedback is inconsistent or delayed. In an extreme case, feedback is inaccurate (a “wicked” learning environment), meaning that the learner, rather than failing to learn the correct relationships, or learning nothing at all, will end up learning incorrect relationships. Expertise can only develop in a kind learning environment. Note that the learning environment is also, by definition, a property of the environment rather than the individual. However, in contrast to the task environment, the learning environment is modifiable, in that it can change for the better or worse.

The third dimension is related to engaging in sufficient structured practice, such as “deliberate practice” (Ericsson et al., 1993), to be able to learn relationships. Learners will need to have high levels of motivation to be able to engage in long periods of structured practice in the context of a kind learning environment. In this sense, this third dimension reflects a property of the individual. A detailed description of deliberate practice is beyond the scope of the current chapter. Stated briefly, deliberate practice can be described as a training method in which the learner is: 1) Given a task exceeding his or her current skill level; 2) Motivated to practice extensively and improve (a generally effortful endeavor); 3) Provided with rapid, comprehensive, and accurate feedback; and 4) Prompted to reflect on the learning experience. Once this goal has been met, the trainee advances to a more difficult task.

The concept of deliberate practice was first identified as the critical factor in achieving expertise in “motor skills”, such as in sports or music performance. The concept has since been extended to more “cognitive domains”, such as chess (Gobet & Campitelli, 2007) and clinical psychology (Miller et al., 2020). One can distinguish two forms of the deliberate-practice hypothesis. In the weak form, extended deliberate practice (e.g., for 10,000 hours) is a necessary cause of expertise. That is, in order to be an expert you need to practice (i.e., it is not possible to become an expert without practice), although other attributes are required. This implies that not everyone can be an expert. In the strong form, practice is both a necessary and sufficient cause of expertise. That is, in order to be an expert, you need to practice, and nothing else is required. The strong form appears to be wrong in many domains of expertise, for example chess and music (Gobet, 2016; Hambrick et al., 2014).

To summarize, to develop true expertise in a particular task the following conditions must be met: 1) There must be reliable relationships between cues and outcomes in the environment that are “learnable”, 2) The learner is in a kind learning environment, and 3) The learner engages in long periods, usually at least ten years, of structured “deliberate practice”. If all these conditions are met, then expertise is likely to develop.

An extreme example of a highly regular task environment, with (typically) a kind learning environment, is chess. Here, all information is available to the two players. There are clear relationships between cues (patterns of chess pieces) and outcomes that can be learned. That is, the outcome is predictable from the cues that a chess player is exposed to while playing. Expertise can potentially develop in such an environment. Generally, the learner receives fast and accurate feedback through play. For example, they learn whether they win or lose the game, and they also learn some information about specific moves (e.g., that a move is weak because it allows the opponent to employ a tactical motive). Of course, human coaches and chess engines are another source of feedback regarding the quality of moves and chess analysis.

An extreme example of an irregular (or chaotic) task environment is a roulette wheel. Here, assuming a balanced wheel, there are no relationships between cues and outcome to be learned. That is, the outcome is not predictable from the myriad of cues that a gambler is exposed to while playing. Expertise cannot of course develop in such an environment, and, absent compelling evidence to the contrary, we would discount claims of expertise on performance on roulette wheels. Note that for roulette wheels, the would-be learner has exposure to lots of rapid and accurate feedback (where the ball lands after each spin), so one can describe it as a kind learning environment. In addition, the would-be learner could invest unlimited amounts of time in “practicing” certain techniques and strategies while performing the task. But the problem is that there are no relationships to learn, and therefore expertise cannot develop. Unfortunately, as is shown with problem gambling “expertise”, spurious patterns and regularities might be found in such environments. They are not predictive of anything, but the hapless gambler believes that they are, and keeps playing until his or her last dollar is lost (Gobet & Schiller, 2014).

Although playing a roulette wheel and problem gamblers may seem like an extreme example, there may be many real-world examples that bear more of a resemblance to the roulette wheel than to chess. In particular, experts’ mid- to long-term political predictions have been shown to be surprisingly poor, and no better than chance performance (Tetlock, 2005). Kahneman (2011) also argues that the performance of stock pickers is rarely much better than chance performance, and does not constitute a true expertise. The poor performance of mid- to long-term predictions is likely mainly due to two factors. First, the environment may be largely unpredictable over a longer timescale. That is, in common with the roulette wheel, there may be few true relationships between cues and outcomes that are learnable over this timescale. Stated more informally, longer-term predictions may be an impossible task. Second, in contrast to the roulette wheel, even if there were some weak relationships that are in principle learnable in certain domains (such as political prediction or stock picking), there would be relatively few

“trials” to learn relationships, meaning that the learning environment is not kind. This makes learning difficult due to a lack of opportunity to engage in a sufficient amount of deliberate practice. For example, it is difficult to make predictions regarding Putin’s actions of the war in Ukraine when there are insufficient data for learning any pertinent relationships. It would be like trying to learn how to master an opening in chess from a review of a handful of games.

So far, we have emphasized chess as a high-validity domain, and roulette wheels as a zero-validity domain, because they reflect two extremes. In reality, many domains likely involve a mix of lower- and higher-validity environments for different tasks. For example, in clinical psychology, Kahneman (2011, p. 242) notes that therapeutic skills (a particular type of task) may develop in high-validity environments, because of the fast and continuous feedback that individuals receive during these activities. In contrast, long-term prediction of clinical outcomes (another type of task), such as anticipating how likely a particular patient is to benefit from a treatment, may involve a low-validity environment, because the learner will be unlikely to get the information required to learn any relationships. Therefore, although beyond the scope of the current chapter, a thorough analysis of each domain should involve separate analyses of different tasks within each domain, as they may have different characteristics.

A broader issue with the development of expertise in many domains is summarized by Gilovich (1993, p. 3), who writes that: “The world does not play fair. Instead of providing us with clear information that would enable us to ‘know’ better, it presents us with messy data that are random, incomplete, unrepresentative, ambiguous, inconsistent, unpalatable, or secondhand.” Stated in terms of the current framework, in many domains learners may be operating in an unkind learning environment, at least for some tasks, and this learning environment does not support efficient learning of whatever relationships are in principle learnable.

So far, we have focused on outlining the framework for understanding trustworthy and untrustworthy experts, but have described expert performance itself in vague terms. In many domains, there is a clear objective measure of performance by which one can identify experts

(Gobet, 2016). For example, chess ratings, such as Elo ratings, capture performance in relation to patterns of results (wins, draws, losses) over a period of time, taking into account the strength of the opponents. In athletics, performance in various events is captured by running times, for example. During the development of expertise, performance on objective measures improves, or becomes more accurate and less variable, or, equivalently, more consistent. In some domains, there are fewer obvious objective markers of performance (such as running times in a 100m race). In those domains, performance is assessed by more subjective measures, such as evaluation by peers. To draw the distinction with the expert performance based on objective measures, the term “respect experts” has been used (Kahneman et al., 2021).

To reiterate, the combination of the properties of the task environment, the kindness of the learning environment, and the extent of structured or deliberate practice will generate different types of expertise (Figure 1). When all three come together, we argue that true expertise is observed. This is located at the top right corner of Figure 1, and is the case for chess masters. Individuals located in this space are “trustworthy subjects”. At other places in Figure 1, we will be dealing with less trustworthy subjects.

Figure 1 about here

---

2. Experts vs Other Experts vs Other Stakeholders

If we assume that an expert has at least some level of trustworthiness, the question arises as to whether we should “follow the expert”. In many domains, this is uncontroversial. For example, if our life depended on finding the best move in a chess position, we would seek advice of the top human player/s. If we consulted multiple top human players, we could take the most popular choice as our move. Alternatively, if we were allowed to consult chess engines, we would use their recommendations. In this scenario, there is limited doubt that we should “follow the expert”, whether derived from human or machine expertise.

However, in many domains, particularly in the social, political, and medical sciences, there are at least four complications. First, as already noted, expertise in these domains may involve less trustworthy expertise, at least for some tasks. Second, two experts working in the same domain may disagree widely. This is of course routinely observed in the legal domain (Mieg, 2001), as described in more detail below. Third, an expert working in one domain may arrive at a different conclusion regarding appropriate actions than an expert working in a different domain. Last, as noted earlier, “there is a growing tension between experts’ recommendations and alternative views – not necessarily grounded on the scientific method –, which appeal to and necessarily involve a different set of norms and values”. How should alternative views, based on different values, be weighed against those of the expert/s? And should this depend on the level of trustworthiness of the expert/s?

As noted above, the legal domain provides compelling examples of where experts can be recruited to argue for different, even opposite, positions. This is particularly apparent in legal systems using Common Law, used for example in the United States and the United Kingdom, which rely heavily on precedents (cases). An expert witness may be selected by the prosecution and so that they will argue in favor of the prosecutor’s case. An expert witness selected by the defense will of course be expected to argue in favor of the defendant. Expert witnesses generally, but not always, have a scientific background. Expert testimonies are delivered in an adversarial setting, with the opportunity for cross-examination from the opposing team. As noted by Mieg (2001), three points can be highlighted. First, as noted above, experts can be expected to contradict each other. Second, experts are not used primarily to establish the “truth”, but to support the case made by the prosecution and/or the defense. Third, the most effective expert witnesses are those who are able to communicate their viewpoint, perhaps with great confidence, rather than those who have the greatest subject matter expertise or those who are most likely to be “correct”.

How, then, do we decide to “follow the expert” if there are two experts with opposite arguments? More broadly, how can we develop trust in these experts? The short answer is that it is probably unwise to put too much trust in experts in the legal context, but that it is easier to trust experts in other domains (such as chess). The financial scandals of the 2000s (the Enron case, and the subprime mortgage crisis) are other examples where trust in experts can be eroded which can make it more difficult to follow the expert (Mieg, 2006).

However, from a broader perspective, Giddens (1990) has taken a different view and has emphasized the merit of “expert systems” which are defined as “systems of technical accomplishment or professional expertise that organize large areas of the material and social environments in which we live today” (1990, p. 27). Expert systems are everywhere and we rarely notice them. For example, for example, if we take a train we must trust that it will transport us effectively, even though we have little to no knowledge of the engineering behind the operation of the transport system. Giddens (1990) argues that in general we have to trust expert systems, and that although problematic counterexamples can be found, in aggregate expert systems are beneficial to individuals and society.

3. Expertise for a New World

In addition to competing viewpoints from other experts, and non-experts and other stakeholders, the application of machine learning and artificial intelligence will provide another input for decision making beyond to those provided by the human experts. Indeed, it is important to consider how machine expertise will interact with human expertise in this context. Another way to think about this is that there are three sources of inputs for decision making: human expert/s (with potentially differing opinions), machine expert/s (with potentially differing “opinions”), and other stakeholders (with potentially differing opinions). As above, we can consider how these three sources of information should be weighed.

One should note that there is long history of research comparing human decision making (some using purported experts) with algorithms, particularly regarding prediction of future

outcomes. As already noted, human experts' prediction of mid- to longer term outcomes have often been shown to be poor, so it is reasonable to ask whether algorithms could do better. In a typical study, a set of predictor variables (e.g., for a personnel selection task, the ratings of candidates on leadership, communication, interpersonal skills, job-related technical skills, motivation) are used to predict a target outcome (subsequent job evaluations of the same people). Expert human judges make their predictions. An algorithm rule (such as multiple regression) uses the same predictors to produce mechanical predictions of the same outcomes. Overall accuracy of human and mechanical predictions is compared. The typical result is that the algorithms do no worse than the human judges, and often perform better (for personnel selection, see Kuncel et al., 2013; for clinical prediction, see Gardner et al., 1996).

For example, in one study in personnel selection (Yu & Kuncel, 2020), doctoral-level psychologists, employed by an international consulting firm to make such predictions, achieved a correlation of .15 with subsequent performance evaluations, whereas multiple regression yields a correlation of .32 (Multiple R) with subsequent performance evaluations. In another prototypical study examining admissions described in Kahneman (2011), subjective impressions of trained professionals were compared against a simple rule in predicting end-of-year grades. Trained counselors predicted end-of-year grades using the following information: 45-min interview; high-school grades; multiple aptitude tests; and a 4-page personal statement. These judgements were compared against a statistical model that incorporated a smaller number of predictors. The regression equation predicted end-of-year grades using only high-school grades and one aptitude test. Once again, the algorithmic judgment performed better than the human judgments.

One may wonder why the subjective impressions of trained professionals perform poorly at these kinds of prediction tasks, usually no better (and often worse) than simple algorithms. One possibility is that the human experts try to be too clever in their evaluation and combine the features in complicated ways that will rarely capture their influence. A second possibility is that

humans are “noisy”, and often give different answers to the same information. If human experts are noisy in this task, we would not expect low-reliability judgements to be good predictors of real-world outcomes.

The examples referred to above, i.e., prediction in personnel selection, prediction of academic performance, and prediction of mental health outcomes, are clearly not zero-validity, because both human and algorithmic prediction was better than chance. That is, there are some relationships between cues and outcomes that are learnable and useful. But overall prediction was not very high, so they can be considered relatively low-validity domains.

A recent example of a study of forecasting in the social sciences is illustrative of these general themes (The Forecasting Collaborative, 2023). Two forecasting tournaments tested the accuracy of predictions of societal change in a number of domains in the social sciences (ideological preferences, political polarization, life satisfaction, sentiment on social media, and gender–career and racial bias). Volunteer forecasting teams comprising of social scientists were provided with historical trend data on the relevant domains, and teams submitted monthly forecasts for a year (Tournament 1: 86 teams, 359 forecasts), with an opportunity to update forecasts based on new data six months later (Tournament 2: 120 teams, 546 forecasts). The main finding was that social scientists’ forecasts were in aggregate no more accurate than those of simple algorithms (e.g., means of past data, linear regressions). Social scientists’ forecasts were also no more accurate than the aggregate forecasts of a sample from the general public (N = 802).

The authors offer a range of possible explanations for the apparent poor predictions of the social scientists. Most important for our current purposes, they note that social systems may be largely chaotic (our word, used in an informal sense) meaning that accurate prediction is a hard task. The authors use the following language to express this idea: “Like other dynamical systems in economics, physics or biology, societal-level processes may also be genuinely stochastic rather than deterministic” (p. 10). Stated in terms of the language used in the chapter,

predictions in the study were poor because the domains assessed may constitute low-validity environments in which both humans and algorithms do not achieve good accuracy.

It should be noted that in addition to the chaotic nature of social systems, the authors offer a number of other possible explanations for the relatively poor accuracy of forecasts of social scientists. First, teams self-selected into the study, and therefore may not be representative of social scientists in general. Second, they may not have been sufficiently motivated to perform well. Third, their knowledge and understanding of effects, often of small size, observed in controlled laboratory settings may not generalize well to real-world settings. Fourth, they may not have received sufficient training in predictive modelling to maximize performance on the task. Last, the study took place during the pandemic, which may have complicated the application of theoretical models. Performance may have been better during a more “normal” period.

The paper did note that there was some role of expertise in the accuracy of forecasts. Specifically, although effect sizes were generally small, scientists were more accurate if they had scientific expertise in a prediction domain, if there were interdisciplinary, and if they used simpler models and based their predictions on historical data. In particular, publication track record on a topic, rather than subjective confidence in domain expertise or confidence in the forecast, contributed to greater accuracy. In addition, the fact that it was possible to identify some predictors of accuracy of forecasts suggests that the domains were low validity, rather than zero-validity. Finally, it is also important to point out that forecasting is only one type of tasks of social scientists may carry out. Of course, they engage in other tasks and have presumably developed expertise in other tasks such as designing and carrying out experimental studies, developing theoretical models, processing and analyzing data, and writing scholarly articles. As noted earlier, it is important to consider the range of tasks that individuals engage in when evaluating expertise,

To summarize, for zero-validity domains, there are no relationships that can be learned, and therefore prediction is not possible for humans or algorithms. For low-validity domains, algorithms tend to perform no worse, or perform better, than human experts. One might wonder how human experts and algorithms fare on high-validity domains, and this is what we consider next.

Once again, we will use chess as an example domain to consider these issues. A computer algorithm, or “chess engine”, Deep Blue, first beat a chess world champion, Garry Kasparov, in 1997 (Campbell et al., 2002). Since that time, chess engines running on a personal computer or handheld computer are so strong that top grandmasters have stopped playing matches against them. Thus, the short answer to the question posed in the previous paragraph is that chess engines are much stronger than the top human players in the domain of chess. We will also provide a longer answer focusing on the evolution of AlphaZero (Silver et al., 2018).

AlphaZero is an artificial intelligence system developed by DeepMind, which uses machine learning to master various games, including chess (Silver et al., 2018). It uses three mechanisms, Monte Carlo tree search, Deep Learning, and reinforcement learning. Rather than searching the tree of possible moves in a systematic way, Monte Carlo tree search generates games by randomly picking moves for the two players. The idea is that, if a move in the current position is better than the alternatives, this move should lead to better results on average, when many such games are played, even though each individual move is selected randomly. With more sophisticated variations of this technique, the choice of moves is biased by previous experience.

Deep Learning consists of adjusting the weights of an artificial neural network, using techniques recently developed (LeCun, Bengio, & Hinton, 2015). AlphaZero uses two networks: the first suggests a move in a given position, and the second evaluates the position as a whole. The program learns by playing a large number of games against itself (“self play”), tuning the weights of its networks using a technique called reinforcement learning. The neural network is

used to evaluate the potential of each possible move, and the Monte Carlo tree search algorithm is used to select the best move based on this evaluation. This technique uses the feedback obtained by the outcome of games to further learn.

Stated simply, in chess AlphaZero starts by training a neural network on the basic rules of the game and then improves its understanding through self-play, where it repeatedly plays games against itself and learns from its mistakes. AlphaZero's approach is highly flexible, allowing it to learn and play new games with little or no prior knowledge of the rules. This makes it a highly sophisticated and powerful AI system, capable of defeating top human players and other computer programs in a variety of games.

There are three especially striking features of AlphaZero that we wish to elaborate on here. First, AlphaZero learns very quickly, and after about 4 hours of training can beat other strong programs (and, presumably, the top human players). Second, AlphaZero plays to a very high standard even if it is not allowed to search positions. The importance of pattern recognition in human experts has been long emphasized (Gobet, 2016) and therefore raises the possibility that AlphaZero is a model for human expertise. Last, the play of AlphaZero led to new insights in chess theory as described in a book, *Game Changer* (Sadler & Regan, 2019). In *Game Changer*, new advances in chess theory are described that have been developed after study of the play of AlphaZero (that had no human tutor). If machine learning can enhance understanding of chess theory, this raises the question as to whether it can enhance understanding of theory in other domains.

To summarize, at least in one high validity domain, machine expertise easily outperforms the top human players, and in one case appears to do so using mechanisms that may be similar to humans. Moreover, machine expertise can provide concrete information that can elevate human expertise. This suggests a model for how human and machine expertise may interact in the future and provides a glimpse of “Expertise in the New World”.

Alternative Perspectives

We now expand the discussion to connect with alternative perspectives on expertise in different domains. First, Von Bertalanffy (1973, p. 39) distinguishes *closed systems*, "i.e. systems which are considered to be isolated from their environment", *with open systems* – all living organisms – which have interactions with their surroundings through input and output flows (e.g. exchanges of energy or information). Traditional physics and chemistry deal with closed systems and highly precise predictions are possible. With other domains – weather forecasting, global warming – one deals with open systems, and the accuracy of prediction drops sharply.

We note that the most regular domains (e.g., chess, tennis, physics, music) are closed systems. As discussed earlier, true expertise can develop in these domains with suitable feedback, practice and motivation. However, with domains that have a high impact for society (e.g., political science and economics), we are often dealing with open systems, which come with more variables, more interactions between processes and variables, and more difficulty learning relationships. The consequence is that it is much harder to learn through practice and study, and therefore that true expertise may not develop. Thus, some of the most trustworthy experts may be found in domains with modest societal importance.

Second, and related, it could also be argued that many domains in which computers display real expertise (e.g., board games, music) are closed systems, and therefore likely to be of little societal import. While the fact that AlphaGo played a brilliant move against human champion Lee Sedol might be fascinating for Go aficionados, it does not help solve pressing societal issues. Furthermore, it has been argued that there is a difference between making a decision, in which computers excel, at least in some domains, and making a choice, which according to authors such as Weizenbaum (1976) and Cisek (1999), is beyond the competence of computers as it entails the question of responsibility (Konigs, 2022). They make the argument that the domains that are important for society require choices, not decisions. However, in

recent years, the boundary between domains that are both important for society and where AI systems perform well has blurred. For example, AI systems are increasingly used to make scientific discoveries in biology (Jumper et al., 2021), medicine (Abd-Alrazaq et al., 2020) and clinical psychology (Morales et al., 2017).

Third, as noted earlier, it is striking that domains that have a large impact on society are often open systems, where predictions are difficult. In domains such as political science, experts are typically valued for their ability to explain events post hoc, and rarely for their ability to draw successful predictions. For example, Putin's annexing of Crimea was predicted by few political scientists but explained after the fact by many. Another interesting observation is that experts in these fields are characterized by use of primarily declarative knowledge (*knowing that*; knowledge of facts that is accessible consciously), whereas experts in closed systems such as chess tend to use primarily procedural knowledge (*knowing how*, knowledge of actions to carry out given a certain situation). The distinction between these two types of knowledge has been discussed in great depth in philosophy (Gobet, 2012; Snowdon, 2004; Winch, 2009), and further discussion is beyond the scope of this chapter. However, the differential use of these two types of knowledge in open and closed systems warrants further analysis.

Last, whilst some AI systems such as AlphaZero display superhuman expertise, it is clearly the case that many do not. The question then arises as to whether humans in general trust AI algorithms. In an intriguing series of online experiments, Krügel et al. (2022) show that people tend to trust AI-powered algorithms, even when they were told that the algorithm is unreliable and therefore should not to be trusted. Participants received advice from an algorithm about the decision to make about ethical dilemmas. They received different types of information about the algorithm. Krügel et al. found that participants trusted the algorithm even when they knew nothing about how it was trained, or even when they were told that it should not be trusted (the algorithm imitated convicted criminals). The results of this study resemble the public's reaction to ChatGPT, which people tend to trust (Choudhury & Shamszare, 2023). Clearly,

educating the public about the reliability and limitations of such AI systems, particularly with respect to their level of expertise, are two important tasks for the industry and society in general.

Conclusion

A central concern of the book is the importance of being able to distinguish trustworthy subjects from untrustworthy individuals. We have addressed this question from the perspective of theory and data from the cognitive psychology of expertise. We have argued that to develop true expertise (i.e., become a trustworthy subject) in a particular task the following conditions must be met: 1) There must be reliable relationships between cues and outcomes in the environment that are “learnable”, 2) The learner is in a kind learning environment, and 3) The learner engages in long periods, usually at least ten years, of structured “deliberate practice”. If all these conditions are met, then expertise is likely to develop. We have also compared human experts with machine experts in a number of domains and found that the latter are often superior, or no worse, than the former. Therefore, human expertise may be enhanced by judicious application of machine expertise, potentially enhancing trustworthiness of human experts.

References

Abd-Alrazaq, A., Alajlani, M., Alhuwail, D., Schneider, J., Al-Kuwari, S., Shah, Z., Hamdi, M., & Househ, M. (2020). Artificial intelligence in the fight against COVID-19: Scoping review. *Journal of Medical Internet Research, 22*, 1–18.

Campbell, M., Hoane, A. J., & Hsu, F. H. (2002). Deep Blue. *Artificial Intelligence*, 134, 57-83.

Choudhury, A. & Shamszare, H. (2023). *Investigating the impact of user trust on adoption and use of ChatGPT: A survey analysis*. 10.2196/preprints.47184.

Cisek, P. (1999). Beyond the computer metaphor: Behaviour as interaction. *Journal of Consciousness Studies, 6* (11-12), 125–142.

Ericsson, K. A., Krampe, R. Th., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review*, 100, 363-406.

Gardner, W., Lidz, C. W., Mulvey, E. P., & Shaw, E. C. (1996). Clinical versus actuarial predictions of violence in patients with mental illnesses. *Journal of Consulting and Clinical Psychology*, 64(3), 602–609

Giddens, A. (1990). *The consequences of modernity*. Cambridge, UK: Polity Press.

Gilovich, T. (1993). *How we know what isn't so*. New York: Free Press.

Gobet, F. (2016). *Understanding expertise: A multidisciplinary approach* (2016). Palgrave MacMillan, UK.

Gobet, F. (2012). Concepts without intuition lose the game: Commentary on Montero and Evans (2011). *Phenomenology and the Cognitive Sciences, 11*, 237-250.

Gobet, F., & Campitelli, G. (2007). The role of domain-specific practice, handedness and starting age in chess. *Developmental Psychology, 43*, 159-172.

Gobet, F., & Schiller, M. (Eds.). (2014). *Problem gambling: Cognition, prevention and treatment*. London: Palgrave.

Hambrick, D. Z., Oswald, F. L., Altmann, E. M., Meinz, E. J., Gobet, F., & Campitelli, G. (2014). Deliberate practice: Is that all it takes to become an expert? *Intelligence, 45*, 34–45.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature, 596*, 583–589.

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.

Kahneman, D., Sibony, O., & Sunstein, C. (2021). *Noise: A flaw in human judgment*. Little, Brown Spark.

Konigs, P. (2022). Artificial intelligence and responsibility gaps: What is the problem? *Ethics and Information Technology, 24* (36), 1–11.

Krügel, S., Ostermaier, A., & Uhl, M. (2022). Zombies in the loop? Humans trust untrustworthy AI-advisors for ethical decisions. *Philosophy & Technology, 35*, 17.

Kuncel, N. R., Klieger, D. M., Connelly, B. S., & Ones, D. S. (2013). Mechanical versus clinical data combination in selection and admissions decisions: A meta-analysis. *Journal of Applied Psychology*, 98(6), 1060–1072.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521, 436-444.

McGrath, T., Kapishnikov, A., Tomašev, N., Pearce, A., Wattenberg, M., Hassabis, D., Kim, B., Paquet, U., & Kramnik, V. (2022). Acquisition of chess knowledge in AlphaZero. *Proceedings of the National Academy of Sciences of the United States of America*, 119(47), e2206625119.

Mieg, H. A. (2001). *The social psychology of expertise: Case studies in research, professional domains, and expert roles*. Mahwah, NJ: Erlbaum.

Mieg, H. A. (2006). Social and sociological factors in the development of expertise. In K. A. Ericsson, N. Charness, P. Feltovich & R. R. Hoffman (Eds.), *Cambridge handbook of expertise and expert performance* (pp. 743-760). Cambridge, UK: Cambridge University Press.

Miller, S. D., Hubble, M. A., & Chow, D. (2020). *Better results: Using deliberate practice to improve therapeutic effectiveness.* American Psychological Association.

Morales, S., Barros, J., Echávarri, O., García, F., Osses, A., Moya, C., Paz Maino, M., Fischman, R., Núñez, C., Szmulewicz, T., & Tomicic, A. (2017). Acute mental discomfort associated with suicide behavior in a clinical sample of patients with affective disorders: Ascertaining critical variables using artificial intelligence tools. *Frontiers in Psychiatry, 8,* 1–16.

Sadler, M., & Regan, N. (2019). *Game changer: AlphaZero's groundbreaking chess strategies and the promise of AI*. Alkmaar, The Netherlands: New In Chess.

Shanteau, J. (1992a). Competence in experts: The role of task characteristics. *Organizational Behavior and Human Decision Processes, 53*, 252-266.

Shanteau, J. (1992b). How much relevant information does an expert use? Is it relevant? *Acta Psychologica, 81*, 75-86.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., . . . Hassabis, D. (2018). A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419), 1140-1144. doi:10.1126/science.aar6404

Snowdon, P. (2004). Knowing how and knowing that: A distinction reconsidered. *Aristotelian Society, 104,* 1-29.

Tetlock, P. E. (2005). *Expert political judgment*. Princeton, NJ: Princeton University Press.

The Forecasting Collaborative (2023). Insights into the accuracy of social scientists' forecasts of societal change. *Nature Human Behavior*.

Von Bertalanffy, L. (1973). *General system theory.* New York: Braziller.

Weizenbaum, J. (1976). *Computer power and human reason*. New York: W.H. Freeman and Company.

Winch, C. (2009). Ryle on knowing how and the possibility of vocational education. *Journal of Applied Philosophy, 26*, 88-101.

Yu, M. C. and Kuncel, N. R. (2020) Pushing the limits for judgmental consistency: Comparing random weighting schemes with expert judgments. *Personnel Assessment and Decisions*: 6(2), Article 2.

Figure 1 Caption

Figure 1 shows the three dimensions to be considered when evaluating expert performance, the validity of the task environment (ranging from zero to high), the learning environment (ranging from unkind to kind), and amount of structured practice. True expertise (shown in green) can only develop when 1) there are reliable relationships between cues and outcomes in the environment that are “learnable”, 2) The learner is in a kind learning environment, and 3) The learner engages in long periods, usually at least ten years, of structured “deliberate practice”.

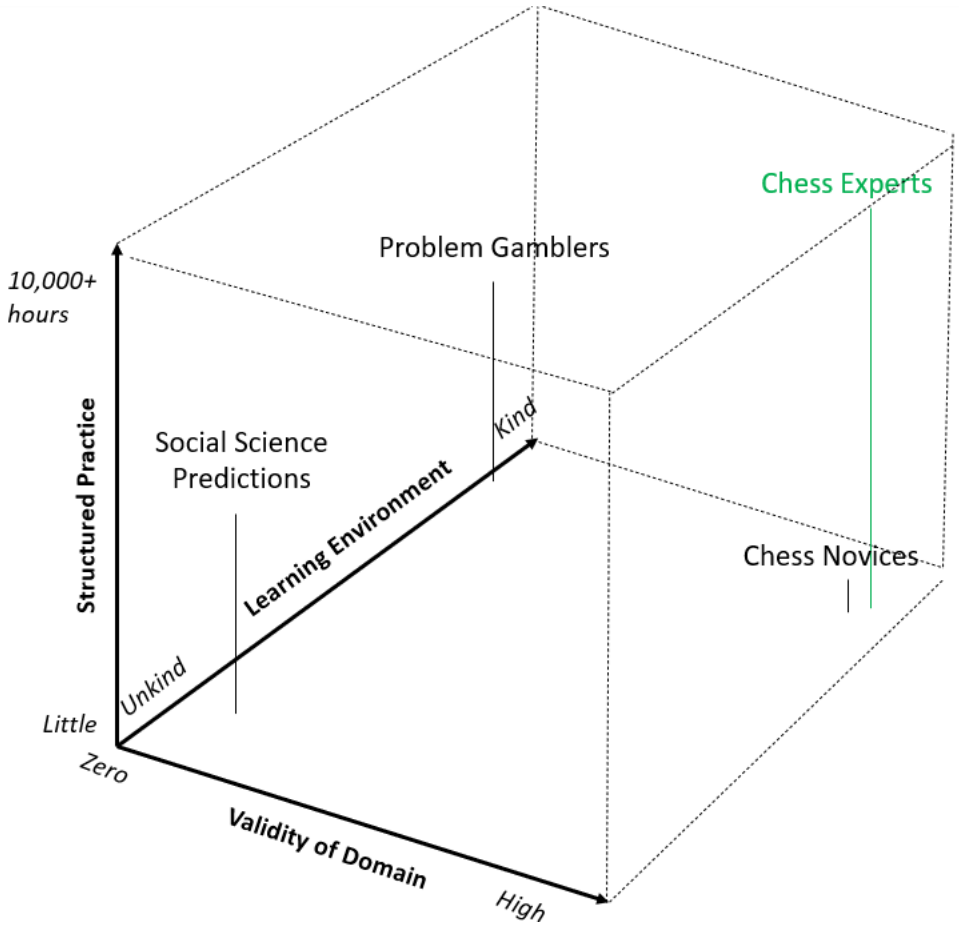
Chess Experts
Problem Gamblers
10,000+
hours
Structured Practice
Social Science
Predictions
Learning Environment
Kind
Chess Novices
Unkind
Little
Zero
Validity of Domain
High

Figure 1