Experts regularly make inconsistent judgments when judging the same case twice. Previous research on expert inconsistency has focused largely on individual or situational factors. Here we focus directly on the cases themselves. First, using a theoretical model, we study how inconsistency and confidence are affected by how strongly experts agree about a case. Second, we empirically test the model’s predictions in two real-world datasets: diagnosticians rating the same mammograms or images of the lower spine twice. Our analyses converge to the same, novel results: cases on which diagnosticians strongly agree are associated with highly confident initial decisions that were unlikely to change—indeed of whether the majority decision was correct or incorrect. Moreover, our results provide clear advice on how to act when faced with two conflicting decisions from a single expert: Choose the more confident decision.

Experts often change their minds, sometimes with profound consequences. For example, a physician might initially classify a mass in a mammogram image as cancerous, but later—when inspecting the same image again—change her mind and classify it as benign. Which diagnosis should the patient rely on? Within-person inconsistency in expert judgments has been widely observed in many domains, including medicine [1–4], clinical psychology [5, 6], neuropsychology [7], finance and management [8], agriculture [9], and weather forecasting [10, 11]. Understanding the causes of expert inconsistency is of key importance, as it may erode societal trust and faces us with uncertainty about the right course of action to take. Here we address two research questions: When do experts—in the absence of any new information—change their decisions? And which decision should individuals rely on? Most studies investigating within-person inconsistency in judgment and decision making have focused either on processes within the individual, such as probabilistic sampling of information [12–14] and hierarchical hypothesis testing [15], or on situational factors, such as time pressure [16]. Furthermore, most previous studies have focused primarily on non-experts and have not linked inconsistency to confidence. How the cases themselves affect inconsistency in a person’s judgments has received comparatively little attention [5, 17]. To the best of our knowledge, no previous study has addressed the interplay between an expert’s confidence, consistency, and the information in a case. This study aims to close this gap in two steps.* First, using a theoretical model [18], we study how within-person inconsistency and confidence in judgments on two-alternative choice tasks are affected by how clearly the information in a case points to either the correct or the incorrect decision. We do this by relating an expert’s within-person consistency (also known as “intrarater agreement”) to the agreement among a population of experts (also known as “interrater agreement”). Next, we empirically test

*Data and code to reproduce the analyses reported in this paper are publicly available at https://osf.io/chgu5. For more information see the Data availability statement at the end of the paper.
the model’s predictions in two real-world datasets: diagnosticians rating the same mammograms [19] or images of the lower spine [20] twice. To preview one major insight not anticipated by previous accounts of expert inconsistency: Cases on which there was clear between-expert agreement were associated with highly confident initial decisions that were unlikely to change—indepedent of whether the experts’ consensus decision was correct or incorrect.

A model linking experts’ inconsistency and confidence to a case’s ambiguity

A fundamental process assumed by many models of cognition, judgment, and decision making is that individuals sample evidence from their environment or memory when making a decision [12–14, 18, 21]. This sampled evidence determines both the decision and the confidence in that decision [18, 21, 22]. The assumption is that an individual samples several pieces of evidence (“cues”) and selects the option for which there is stronger evidence. The more clearly the evidence points to the option, the more confident the individual will be in the accuracy of their decision. In particular, making a second decision about the same case is seen as equivalent to drawing a second sample of evidence. Because the sampling process is probabilistic, the evidence in the second sample can differ from that in the first sample—as can the corresponding decision (e.g., “cancer” vs. “no cancer”) and the individual’s confidence in that decision.

But how does inconsistency in repeated judgments relate to confidence and the information available in the judged case? To theoretically investigate this question, we used a simple model [18] embodying the assumptions outlined above to derive qualitative predictions about the relationship between an expert’s inconsistency, confidence, and case ambiguity (i.e., how clearly the information contained in the case pointed to one decision or the other). In the Discussion, we show that relaxing the model’s assumptions would not change the central results and argue that the predictions made would also emerge from other models of judgment and decision making, such as evidence accumulation models [e.g., 21, 23]. The Self-Consistency Model (SCM) [18] assumes that a decision maker facing a two-alternative choice task samples a fixed, odd number of evidence (“cues”) and chooses the option favored by more cues (i.e., decides between two options using majority voting among cues). Given a probability $p$ of sampling a cue that indicates the correct option (say, “cancer”) and assuming that cues are sampled independently, the probability $P$ of making a correct decision thus follows from the binomial distribution:

\[ P(p, n) = \sum_{h=m}^{n} \binom{n}{h} \cdot p^h (1 - p)^{n-h}, \quad (1) \]

where $m = \frac{n+1}{2}$ (i.e., the minimum number of cues necessary to decide in favor of the correct option). The probability $I$ of making two decisions that are inconsistent is:

\[ I = P(1 - P) + (1 - P)P = 2P(1 - P), \quad (2) \]
which is maximal \( (I = 0.5) \) for choices at chance level \( (P = 0.5) \) and thus by extension for cases that are maximally ambiguous \( (p = 0.5) \), that is, when every sampled cue is equally likely to point to the correct or to the incorrect decision. Conversely, inconsistency is minimal \( (I = 0) \) for perfectly correct \( (P = 1) \) and “perfectly” incorrect \( (P = 0) \) decisions (Figure 1A), that is, when every sampled cue points either to the correct decision \( (p = 1; \text{perfectly “kind” cases}) \) or to the incorrect decision \( (p = 0; \text{perfectly “wicked” cases}) \). Henceforth, we label cases with \( P > 0.5 \) as “kind”, and cases with \( P < 0.5 \) as “wicked”. Thus, in the SCM, within-expert inconsistency increases the closer a cases’s \( p \) is to a fair coin flip.†

The SCM provides a simple, elegant link between the consistency of a single expert’s repeated judgments on a case and between-expert agreement on the same case. For simplicity, let us assume that all experts sample the same number of cues (i.e., share a common \( n \)) and that for any particular case and cue those experts have the same probability \( p \) of sampling a cue that points to the correct answer. Although \( p \) is not directly observable, according to the SCM, the expected proportion of correct decisions \( E(P_i(p_i)) \) for case \( i \) among a population of identical experts is monotonically related to \( p_i \). Empirically, the sample proportion of correct decisions among experts for case \( i \), \( \hat{P}_i \), can be used as a proxy for ordering cases according to their \( p_i \).‡ Thus, we can use the disagreement among experts (i.e., how close \( \hat{P}_i \) is to 0.5) as an indicator of a case’s ambiguity (i.e., how close \( p_i \) is to 0.5).

The SCM further stipulates that the confidence in a decision increases with the proportion of cues pointing to that decision. In particular, the SCM assumes that confidence \( \hat{C} \) is the complement of the sample standard deviation, which depends solely on the proportion of cues pointing to the chosen option (\( \hat{p} \) for correct decisions and \( 1 - \hat{p} \) for incorrect decisions):

\[
\hat{C} = 1 - \sqrt{\hat{p}(1 - \hat{p})}.
\]

Because \( \hat{p} = E(p) \), it follows that confidence is highest for \( p = 1 \) and \( p = 0 \) (\( \hat{C} = 1 \)) and lowest for \( p = 0.5 \) (\( \hat{C} = 0.5 \); Figure 1B)—mirroring the results for an individual expert’s inconsistency (see eq. 2 and Figure 1A). Given that inconsistency \( I \) increases and confidence \( \hat{C} \) decreases with increasing case ambiguity, it follows that confidence and inconsistency are negatively related (Figure 1C).

But which decision should people confronted with two inconsistent, conflicting decisions rely on? According to the maximum-confidence slating (MCS) algorithm (henceforth “confidence

†More formally, a case’s ambiguity is a monotonically decreasing function of \( |p - 0.5| \) (i.e., how close a case’s \( p \) is to a fair coin flip).

‡Because equation 1 applies to majority voting over either cues or individuals, we can use Condorcet’s Jury Theorem [26, 27] to gain insights into how \( P_i \) and \( p_i \) relate for \( n \geq 3 \). For example, for \( p_i > 0.5 \rightarrow P_i > p_i \); conversely, for \( p_i < 0.5 \rightarrow P_i < p_i \). Thus, if we assume that experts sample more than one cue, \( P_i \) will be a more extreme version of \( p_i \). Particularly, for any \( n \geq 3 \), \( P_i \) and \( p_i \) are identically ordered across a set of cases.
rule”) [25], they should adopt the more confident decision [28, 29]. The SCM predicts that confidence will be positively correlated with the probability of making a correct decision for $p > 0.5$, but negatively correlated for $p < 0.5$ (Figure 1B).³

In sum, the SCM predicts:

1. The higher a case’s ambiguity (indexed by disagreement among experts’ initial judgments), the more likely an individual expert will be inconsistent when judging the same case again (Figure 1A).

2. The higher a case’s ambiguity (indexed by disagreement among experts’ initial judgments), the less confident an individual expert will be in their initial judgment (Figure 1B).

3. The less confident an individual expert is in their initial judgment, the more likely they will be to change it when judging the same case again (Figure 1C).

4. Considering only cases in which an individual expert changed their judgment: Relative to sticking with the initial judgment, using the confidence rule (i.e., selecting the more confident judgment) improves accuracy for kind cases (right part of 1B) but worsens it for wicked cases (left part of 1B).

However, these predictions depend on the SCM’s [18] strong assumptions about an expert’s judgment process and our additional assumption of complete homogeneity among experts and cues. Specifically, the model assumes that all experts sample the same number of $n$ cues and that, for any particular case and cue, those experts have the same probability $p$ of sampling a cue that points to the correct answer. These assumptions are unlikely to hold for actual expert judgments. Thus, to empirically test the model’s predictions and assess how much insight it provides into the judgments of real experts, we drew on two real-world high-stakes expert datasets: diagnosticians rating mammograms [19] and X-rays of the lower spine [20] twice.

Empirical analyses: Expert inconsistency in medical diagnostics

Dataset 1: Radiologists diagnosing mammograms [19] Dataset 1 consists of repeated judgments of the same mammograms. The data were collected to study the effect of time spent viewing and confidence on diagnostic accuracy in mammography screening [19]. Of the 572 radiologists

³More specifically, equation 1 shows that, for $p > 0.5$, any level of confidence is more likely to be observed under the correct than the incorrect decision (and vice versa for $p < 0.5$). To see why, consider that, in equation 1, $\bar{p} = \frac{l}{n}$. When $p > 0.5$, it follows that $p^h > (1 - p)^{n-h}$ and thus the event that a majority of cues point to the correct decision ($h$) is more likely than the event that the same-sized majority of cues point to the incorrect decision ($n - h$). For $p < 0.5$, we obtain the opposite result.
invited to participate, 102 completed both phases of the study. The mammograms used were randomly selected from screening examinations of women aged 40–69 years old. The correct diagnosis (cancerous or non-cancerous) for each mammogram was available from follow-up research. In phase 1, each radiologist was randomly assigned to one of four test sets of 109 mammograms. The radiologists were instructed to interpret the cases as they would in clinical practice. They were informed that the overall cancer rate in their test set was higher than that found in a screened population, but they were not informed of the specific prevalence of cancer cases in their test set. When viewing each case, radiologists were prompted to identify the most significant breast abnormality and to decide whether the patient should be recalled for additional workup. The decision to recall constituted a positive test result. Additionally, participants provided a confidence judgment for each assessment (“not at all confident,” “not very confident,” “neutral,” “confident,” or “very confident”). Radiologists used either a home or work computer or a laptop provided by the study to complete the task. After an interval of 3–9 months, radiologists were re-invited to rate a second set of 110 mammograms, following an identical procedure. Unknown to the participants, a proportion of the cases presented in this phase 2 were the same as in phase 1. Overall, 58 cases were rated twice by 55 radiologists; of those 58 cases, 46 were rated twice by another 47 radiologists, resulting in 5,352 repeated ratings. All repeated mammograms were non-cancer cases (i.e., from women who were cancer-free for at least 2 years after the mammography). See [19] for more details. Across all repeated cases, the median accuracy (proportion correct) was 0.72 for phase 1 and 0.68 for phase 2.

**Dataset 2: Physicians diagnosing X-rays of the lumbosacral spine [20]** Dataset 2 consists of repeated judgments of the same X-rays of the lumbosacral spine. These data were collected to
study the diagnostic accuracy of radiologists and chiropractors (total $N = 13$) reading lumbosacral radiographs [20]. Five chiropractors, three chiropractic radiologists, and five medical radiologists participated in the study. Their professional experience ranged from 3 to 21 years. For the study, 300 X-rays of the lumbosacral spine of adult patients were selected from a hospital database. The correct diagnosis for each X-ray was based on clinical findings, if present lab data, and/or MR/CT imaging. The selected X-rays overrepresented “significant abnormalities” (17% of cases), including infections ($n = 7$), malignancies ($n = 15$), fractures ($n = 8$), inflammatory spondylitis ($n = 6$), and spondylolysis ($n = 14$). The set of X-rays was presented in a random order. For each X-ray, the physician evaluated whether a significant abnormality was present (in which case immediate referral to a hospital was required) and gave a confidence rating on a two-point scale (1–2). Three months later, all participants assessed all 300 X-rays again, resulting in 3,900 repeated assessments. See [20] for more details. Across all cases, the median accuracy (proportion correct) was 0.86 in phase 1 and 0.91 in phase 2.

**Statistical analyses.** We ran a series of Bayesian mixed-level regression models using the R package *brms* and its default priors [30]. The models all included group-level intercepts for individuals and cases (“random intercepts”). See Supplementary Material for detailed model descriptions and results (see Supplementary Table S1).

Note that each of the 300 X-rays in the spine dataset was rated by just 13 experts. Consequently, the estimates for both proportion correct $\hat{P}_i$ and inconsistency $\hat{I}_i$ are noisy. In the mammography dataset, in contrast, up to 102 radiologists rated 58 distinct mammograms, allowing the characteristics of the cases to be estimated more reliably. To render our classification of cases in the spine dataset as kind versus wicked more reliable, we defined—in both datasets—kind cases as $\hat{P}_i > 0.6$ and wicked cases as $\hat{P}_i < 0.4$. We thus excluded cases where $0.4 < \hat{P}_i < 0.6$ in model M7 (Supplementary Table S1); those cases were retained in all other analyses and all figures (except Figure 3).

**Results**

Experts’ average performance was better than chance in both studies, but substantially better in the spine dataset (see Supplementary Figure S1). Furthermore, the proportion of wicked cases was lower in the spine dataset. As a consequence, predictions 1, 2, and 4 can be assessed with higher precision for kind than for wicked cases—especially in the spine dataset.

**Prediction 1: The higher a case’s ambiguity, the more likely an individual expert will be inconsistent when judging the same case again.** The results presented in Figure 2A/B corroborate the first prediction: The higher a case’s ambiguity (indexed by disagreement among experts’ initial diagnoses), the more likely it was that an individual expert was inconsistent when judging the same case again—irrespective of whether the between-expert agreement was correct or not: In the mammography dataset (Figure 2A), the more strongly experts initially favored either the
Figure 2: Empirical results on the relationship between the proportion of experts who made a correct diagnosis ($\hat{P}$), inconsistency ($\hat{I}$; probability of making an inconsistent diagnosis), and mean confidence ($\hat{C}$) in the two datasets.  

(A & B) Inconsistency as a function of the proportion of experts who made a correct diagnosis.  

(C & D) Mean confidence as a function of the proportion of experts who made a correct diagnosis.  

(E & F) Inconsistency as a function of mean confidence. Each dot represents one case and its coordinates represent $\hat{P}$ and $\hat{C}$ from initial diagnoses. The solid curves are LOESS smooths. The dashed curves show the smooths when using $\hat{P}$ and $\hat{C}$ from the second diagnoses (the corresponding dot are not shown). Panels B, D, and F employ jittering to avoid overplotting.
correct or incorrect diagnosis, the more consistent their individual responses. In the spine dataset, the same pattern emerged for kind cases (Figure 2B); the results for the few wicked items are also consistent with the prediction, but less conclusive (Figure 2B). Regression model M2 (Supplementary Table S1) shows a negative quadratic term in both datasets, supporting the visual impression from Figure 2A/B. Moreover, the regression analyses show that—prior to accounting for a case’s ambiguity—the cases differed more strongly in how inconsistently they were diagnosed than the experts differed in how inconsistently they diagnosed the cases (see model M1 in Supplementary Table S1). This finding further underlines the relevance of case ambiguity in explaining variation in expert consistency.

Prediction 2: The higher a case’s ambiguity, the less confident an expert will be in their initial judgment. The results presented in Figure 2C/D are in line with the second prediction: The higher a case’s ambiguity (indexed by disagreement among experts’ initial diagnoses), the less confident experts were in their initial diagnoses—again, irrespective of whether the between-expert agreement was correct or not. Regression model M4 (Supplementary Table S1) shows a clearly positive quadratic term in both datasets, supporting the visual impression from Figure 2C/D.

Prediction 3: The less confident an expert is in their initial judgment, the more likely they will be to change it when judging the same case again. The results presented in Figure 2E/F corroborate the third prediction. Regression model M5 (Supplementary Table S1) shows a clearly negative linear term in both datasets, substantiating the visual impression from Figure 2E/F.

Prediction 4: When an expert changes their judgment, using the confidence rule improves accuracy for kind cases but worsens it for wicked cases. Focusing on cases in which an expert’s diagnoses were inconsistent, we found that, relative to sticking with the initial diagnosis, using the confidence rule (i.e., selecting the more confident diagnosis) improves accuracy for kind items. For wicked items (i.e., cases where the majority of initial diagnoses were incorrect), in contrast, results were mixed. The proportion of correct diagnoses in the mammography dataset was similar across first and second judgments. In contrast, second diagnoses in the spine dataset were more accurate than first diagnoses (see Supplementary Figure S1). This latter result suggests that—unless second spine diagnoses are also, on average, sufficiently more confident—the confidence rule is unlikely to outperform the second diagnosis. As Figure 3 shows, for kind cases, using the confidence rule resulted in higher accuracy than either the first or second diagnosis in the mammography dataset. In the spine dataset, this applied only to the first diagnoses. The results for wicked cases were less consistent with our fourth prediction: using the confidence rule did not worsen accuracy. Comparing the confidence rule in both datasets against the first and second diagnosis did not reveal clear differences. Regression model M7 supports these observations.
Figure 3: Empirical results comparing the accuracy of using the confidence rule to the accuracy of first (upper panel) and second (lower panel) diagnoses for cases where experts were inconsistent, separately for wicked and kind cases and the two datasets (mammography and lumbosacral spine). Positive values on the y-axes indicate that the confidence rule outperformed the first or second diagnosis. Cases are shown as horizontally jittered dots; the size of a dot indicates the number of experts contributing to that case. The distributions are summarized by boxplots to the right of the dots. The point and linerange to the left of the dots indicate the median and 95% credible interval, respectively, of the posterior distribution of the expected improvement (according to model 7; see supplementary information).

Discussion

When do experts change their mind? Previous research on within-person inconsistency has focused largely on individual factors (e.g., [12–14]) or situational factors (e.g., [16]). Here we focused directly on the cases themselves. First, using the Self-Consistency Model (SCM) [18], we studied how inconsistency and confidence in judgments on two-alternative choice tasks are affected by how clearly the information in a case points to either the correct or the incorrect decision (a case’s ambiguity, indexed by disagreement among experts’ initial diagnoses). Next, we confirmed three of the model’s four key predictions in two real-world datasets: diagnosticians rating the same mammograms or images of the lower spine twice. We found that the higher a case’s ambiguity, the more likely experts will be to make inconsistent decisions (prediction 1), and the less confident they will be in their initial diagnosis (prediction 2)—irrespective of whether the between-expert consensus opinion was correct or not. The first two results imply that the more confident an expert is in their
initial diagnosis, the less likely they will be to change their diagnosis when judging the same case again (prediction 3), irrespective of whether the between-expert consensus opinion was correct or not. Taken together, these first three results imply that a highly confident or consistent diagnosis is, first and foremost, an indicator for the strength of agreement among experts. It can only be an indicator of accuracy when most cases in the domain of interest are kind. Finally, when an expert’s two diagnoses were inconsistent, using the confidence rule (i.e., selecting the more confident diagnosis) improved accuracy relative to sticking with the initial diagnosis [31]. This fourth prediction was empirically corroborated for kind cases but the opposite result predicted for wicked cases was only partially corroborated—although other results for wicked cases were consistent with predictions 1 and 2 (see the left sides of panels A–D in Figure 2). These mixed findings might be the result of systematic differences in accuracy and confidence judgments between first and second diagnoses, especially for the spine dataset. For example, second spine diagnoses were more accurate than first ones (see Supplementary Figure S1). Future research should explore the implications of such systematic differences. Importantly, however, model M6 (Supplementary Table S1) shows that summarizing across all cases the confidence rule outperformed both first and second diagnoses in the mammography dataset and first, but not second, diagnoses in the spine dataset. Because decision makers cannot tell in advance whether a particular case is kind or wicked [32], using the confidence rule has clear practical merit [31]. Our results thus offer the following advice: Unless you suspect that experts perform worse than chance in the domain of interest, rely on the more confident of two conflicting judgments from an individual expert.

Previous research has primarily studied within-person inconsistency from an internal individual perspective, which explains inconsistency as a consequence of unreliable processing of information. Studies reported, for example, that an individual’s judgments become less reliable as the amount of available information increases [33, 34] because their capacity to process that information decreases [35, 36]. Other studies found that unreliability in judgment can often be attributed to a lack of cognitive control—how acquired knowledge is used—rather than a lack of knowledge [37]. Harvey [17] reviewed further reasons for inconsistency in decisions, such as overload in working memory, learning correlations instead of functions, or reproducing noise in the data. In contrast, external task-related factors contributing to experts’ consistency, such as predictability of the environment (i.e., the degree to which cues allow the outcome to be predicted) have been rarely studied. A set of studies has shown that individuals make less consistent judgments as a task becomes less predictable [17, 38, 39]. Our results are largely consistent with these findings. However, none of these previous studies made the connection between an individual’s confidence, consistency, and the ambiguity of a case. In the following, we discuss three contributions that our perspective on within-expert inconsistency makes.

First, to reduce inconsistency and thus improve accuracy, previous perspectives suggest the use of interventions and methods that improve the overall reliability of information processing, such as reducing the amount of information presented [40], decomposing a complex task into smaller subtasks [41, 42], combining an individual’s repeated judgments [43–45], or requiring individuals to justify their judgments [46, 47]. Our perspective suggests a different, but complemen-
tary approach to improving accuracy, namely, encouraging experts to make a second assessment whenever their initial decision was low in confidence and to then apply the confidence rule across the two decisions. The rationale for this approach is twofold. First, in general, one can expect that experts will perform better than chance and that the confidence rule will therefore improve accuracy relative to sticking to the initial decision. Second, there is little benefit in judging cases that were initially diagnosed with high confidence again; such confident decisions are unlikely to change and the confidence rule will therefore not change the final decision. For example, radiologists often evaluate a set of cases and could record their diagnosis along with other information, such as confidence, in a software application. The software could be programmed to present some cases (e.g., low-confidence cases) again. Then the confidence rule could be implemented in different ways. In one implementation, the radiologist would be shown her inconsistent judgments and asked to make a final decision (e.g., by applying the confidence rule). Another possibility would be for the software to automatically implement the confidence rule. If a software-driven approach is not feasible, research on the wisdom of the inner crowd [43, 44] suggests that prompting individuals to reconsider their initial decision and make a new decision can elicit diverse knowledge that can in turn lead to diverse judgments. Aggregating such diverse judgments can cancel out error, resulting in a more accurate judgment than the initial one. In a similar way, radiologists could use this approach for low-confidence diagnoses. Our results, combined with the research on the wisdom of the inner crowd [43, 44], suggest that it is beneficial to deliberately reconsider low-confidence decisions, and then, if two decisions are inconsistent, to apply the confidence rule.

Second, previous accounts of expert inconsistency explicitly or implicitly assume that accuracy increases as the consistency of judgments increases. In stark contrast to this assumption, our results show that this relationship is mirrored at chance level: For cases that experts tend to judge incorrectly, individual expert consistency starts to increase again the more experts agree on the incorrect diagnosis. Furthermore, our results show that confidence tracks consistency, but because confidence tracks the ambiguity of a case and not accuracy per se [18], its ability to predict accuracy and consistency strongly depends on the environment, that is, the distribution of ambiguity across cases [28]. If there are only kind cases (i.e., cues tend to point to the correction decision), confidence strongly predicts that a diagnosis is accurate and will not change. The more wicked cases there are (i.e., cues tend to point to the wrong decision), the more these relations dilute. In the extreme—and hopefully only hypothetical—case of a domain dominated by wicked cases (in which experts, on average, tend to make wrong decisions), the relations reverse: Experts’ confidence is then negatively related to accuracy, but still positively related to consistency—and being consistent in a wicked environment means confidently sticking to the wrong diagnosis.

Third, previous accounts have focused on differences in consistency among experts or in different task conditions (e.g., time pressure). Our perspective predicts that the cases themselves can differ markedly in how consistently they are diagnosed by any expert. Importantly, as our results have shown, these consistency differences among cases can be even larger than those observed among experts and can be explained to a large degree by a case’s level of ambiguity.
We used the Self-Consistency Model (SCM) [18] to gain insights into when an expert is inconsistent and what to do as a decision maker when faced with inconsistent advice from the same expert. Our implementation of the SCM assumes that all experts sample the same number of \( n \) cues and that for any case and cue those experts have the same probability \( p \) of sampling a cue that points to the correct answer. We argue that relaxing these assumptions will, in general, not change the four key predictions; it will affect the functional form with which the probability of a correct decision, \( P \), depends on \( p \) and \( n \), but the qualitative implications of the distinction between kind cases \( (p > 0.5) \) and wicked cases \( (p < 0.5) \) should remain unchanged. In addressing this point, we can benefit from the fact that the SCM’s decision process amounts to majority voting among cues; we can therefore apply insights from more general research on majority voting. For example, similar conclusions follow if, within an expert, the cues’ probabilities \( p_i \) are not identical but symmetrically distributed around \( p \) [27], or if the retrieval of cues is not independent (e.g., retrieving cues pointing to one option increases the likelihood that further cues point to that same option [27, 48, 49]). As another example, under very general conditions, as the number of cues retrieved, \( n \), increases, the probability of a correct decision, \( P \), will increase for kind cases \( (p > 0.5) \) and decrease for wicked cases \( (p < 0.5) \) [27, 48, 49]). As a consequence, when everything else is kept constant, consistency should increase as more cues are retrieved (see eq. 2); the variation in cases’ ambiguity will be most pronounced for small \( n \)s, whereas for large \( n \)s all cases will be clearly diagnosed either correctly or incorrectly. Furthermore, assuming that experts sample different numbers of cues implies that, for the same case, experts with larger \( n \)s will be more consistent than experts with smaller \( n \)s. Importantly, experts with larger \( n \)s will only be more accurate for kind cases and will be less accurate for wicked cases.

Here we used the SCM [18] as a simple model linking accuracy, confidence, and consistency, but we argue that a broad family of models makes qualitatively similar predictions. To the best of our knowledge, the ability of these models to provide insights into expert consistency and the role of wicked cases has not yet been explored. For example, in the diffusion decision model [23], a prominent example from the family of evidence accumulation models, case ambiguity is reflected in the drift rate, which represents the average speed with which an individual accumulates evidence that stochastically drifts toward one of two decision boundaries (e.g., correct vs. incorrect answer). With everything else kept constant, a reduction in the drift toward zero implies increasingly more ambiguous cases, which are predicted to be associated with lower accuracy, longer response times, and lower confidence [21, 23]. Drift rates below zero represent wicked cases, where the evidence tends to accumulate to the wrong decision boundary. Notably, an increasingly negative drift rate corresponds to increasingly less ambiguous and more wicked cases, which are predicted to be associated with lower accuracy, but shorter response times and higher confidence—thus qualitatively mirroring the predictions from the SCM. More generally, any model that assumes or implies the following two relations should predict qualitatively similar results: the more clearly the relevant information points to an option, the more likely a particular decision becomes and the more confidently it will be rendered. We would argue that those two relations are fundamental to many, if not most, psychological and normative models of decision making. The crux of the matter is how a particular model operationalizes the notion of how clearly information points to the chosen
decision. Future research should map out which decision environments result from the interaction between a decision strategy and the statistical structure among cues and the criterion to be inferred. For example, certain decision strategies, such as lexicographic rules (e.g., take the best or tallying [50, 51]) or exemplar-based reasoning [52, 53], might not yield the same relationship between confidence, inconsistency, and case ambiguity.

Inconsistency in individual experts’ judgments is a common finding in many domains, including medicine, finance and management, and weather forecasting. Most previous studies have investigated inconsistency from an individual or situational perspective, leading to methods to improve information processing within individuals. Here we connected—theoretically and empirically—the ambiguity of the case with the confidence and inconsistency of the expert. For experts themselves as well as individuals confronted with inconsistent judgments from a single expert, we advise the following: Unless there is reason to believe that the expert performs below chance, rely on the more confident judgment.

References


### Acknowledgements
We thank Susannah Goss for editing the manuscript, and Annemarie de Zoete and Raymond Ostelo for sharing the lumbosacral spine dataset. We thank Jose Cayere, Amy Buzby, and the American College of Radiology for their technical assistance in developing and supporting the implementation of the mammography test sets; the expert radiologists Larry Bassett, Barbara Monsees, Ed Sickles, and Matthew Wallis; and the participating women, facilities, and radiologists for the data they provided. The Breast Cancer Surveillance Consortium (BCSC) investigators are listed at [https://www.bcsc-research.org/contact](https://www.bcsc-research.org/contact). This work was supported by the American Cancer Society using a donation from the Longaberger Companys Horizon of Hope Campaign (Grants SIRSG-07-271, SIRSG-07-272, SIRSG-07-273, SIRSG-07-274-01, SIRSG-07-275, SIRSG-06-281, SIRSG-09-270-01, SIRSG-09-271-01, and SIRSG-06-290-04); by the Breast Cancer Stamp Fund; and by the National Cancer Institute Breast Cancer Surveillance Consortium (Grant HHSN261201100031C). The collection of cancer and vital status data used in this study was supported, in part, by several state public health departments and cancer registries throughout the United States. A full description of these sources is provided at [https://www.bcsc-research.org/](https://www.bcsc-research.org/).

### Author contributions
All authors designed research; S.M.H. derived analytical model results; A.L. and S.M.H. analyzed data; A.L. and S.M.H. wrote the paper with input from all other authors.

### Data availability
The spine dataset analyzed in this study is available on the open science framework ([https://osf.io/chgu5](https://osf.io/chgu5)). The mammography data that support the findings of this study are available from [https://www.bcsc-research.org/contact](https://www.bcsc-research.org/contact). However, restrictions apply to the availability.
of these data, which were used under license for the current study, and so are not publicly available. The code to analyze both datasets can be found here: https://osf.io/chgu5.

**Competing Interests** The authors declare that they have no competing financial interests.

**Correspondence** Correspondence and requests for materials should be addressed to A.L. (email: litvinova@mpib-berlin.mpg.de).
Regression Analyses

Methods

We ran a series of Bayesian mixed-level regression models (see Table S1) using the R package brms (version 2.9.0) and its default priors (Bürkner, 2016). The models all included group-level intercepts for individuals and cases (“random intercepts”). Four chains, each with 6,000 samples (and thinning = 4), were run. The first 2,000 samples were discarded as warm up; thus, a total of 4,000 samples were obtained per model. The MCMC diagnostics did not indicate any problems (see supplementary code and outputs at https://osf.io/chgu5).

The three models of inconsistency (M1, M2, and M4) are logistic regression models; their parameters thus indicate (changes in) log odds. The two models of confidence (M3 and M5) are linear models (i.e., identity link).

\((\hat{P} - 0.5)\) and \((\hat{P} - 0.5)^2\) in models M2 and M4 are the linear and quadratic polynomial contrasts of the 0.5-centered proportion of correct diagnoses per case; this means that the intercept in those models predicts the value of the dependent variable for a maximally ambiguous case (\(\hat{P} = 0.5\); because for \(\hat{P} = 0.5\), \((\hat{P} - 0.5) = (\hat{P} - 0.5)^2 = 0\). \((C - 1)\) in model M5 is the linear effect of confidence, re-coded so that the intercept indicates the inconsistency at the lowest confidence level in both datasets (i.e., \(C = 1\); because for \(C = 1\), \((C - 1) = 0\) correspond to the lowest possible confidence rating).

The two models for the confidence rule (M6 and M7) are logistic regression models; their parameters thus indicate (changes in) log odds. These two models were estimated on
Bayesian mixed-level regression models testing predictions 1–4 in the mammography and lumbosacral spine datasets. Posterior distributions of parameters are summarized by their posterior median (Estimate) and 95% credible interval (95% CI). sd(expert) and sd(case) show the standard deviations of the group-level distribution of the intercept for experts and cases, respectively. See main text for more details (incl. coding of variables and interpretation of effects).

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Mammography</th>
<th></th>
<th></th>
<th>Lumbosacral spine</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>M1: Inconsistency (intercept-only model)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>−1.52</td>
<td>−1.76</td>
<td>−1.28</td>
<td>−2.31</td>
<td>−2.72</td>
</tr>
<tr>
<td>sd(expert)</td>
<td>0.47</td>
<td>0.37</td>
<td>0.58</td>
<td>0.62</td>
<td>0.41</td>
</tr>
<tr>
<td>sd(case)</td>
<td>0.77</td>
<td>0.61</td>
<td>0.98</td>
<td>0.95</td>
<td>0.78</td>
</tr>
<tr>
<td>M2: Inconsistency vs. case ambiguity (Prediction 1): $I \sim (\hat{P} - 0.5) + (\hat{P} - 0.5)^2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>−1.50</td>
<td>−1.63</td>
<td>−1.37</td>
<td>−2.39</td>
<td>−2.78</td>
</tr>
<tr>
<td>($\hat{P} - 0.5$)</td>
<td>−1.79</td>
<td>−8.74</td>
<td>5.44</td>
<td>−56.86</td>
<td>−63.49</td>
</tr>
<tr>
<td>($\hat{P} - 0.5)^2$</td>
<td>−49.06</td>
<td>−56.53</td>
<td>−42.52</td>
<td>−27.39</td>
<td>−33.49</td>
</tr>
<tr>
<td>sd(expert)</td>
<td>0.47</td>
<td>0.37</td>
<td>0.58</td>
<td>0.64</td>
<td>0.41</td>
</tr>
<tr>
<td>sd(case)</td>
<td>0.16</td>
<td>0.02</td>
<td>0.30</td>
<td>0.08</td>
<td>0.00</td>
</tr>
<tr>
<td>M3: Confidence (intercept-only model)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>3.73</td>
<td>3.62</td>
<td>3.84</td>
<td>1.62</td>
<td>1.45</td>
</tr>
<tr>
<td>sd(expert)</td>
<td>0.46</td>
<td>0.40</td>
<td>0.53</td>
<td>0.28</td>
<td>0.20</td>
</tr>
<tr>
<td>sd(case)</td>
<td>0.25</td>
<td>0.21</td>
<td>0.31</td>
<td>0.16</td>
<td>0.15</td>
</tr>
<tr>
<td>M4: Confidence vs. case ambiguity (Prediction 2): $C \sim (\hat{P} - 0.5) + (\hat{P} - 0.5)^2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>3.73</td>
<td>3.63</td>
<td>3.83</td>
<td>1.62</td>
<td>1.45</td>
</tr>
<tr>
<td>($\hat{P} - 0.5$)</td>
<td>2.91</td>
<td>−0.34</td>
<td>6.19</td>
<td>5.45</td>
<td>4.32</td>
</tr>
<tr>
<td>($\hat{P} - 0.5)^2$</td>
<td>13.84</td>
<td>10.46</td>
<td>17.07</td>
<td>4.19</td>
<td>3.03</td>
</tr>
<tr>
<td>sd(expert)</td>
<td>0.46</td>
<td>0.40</td>
<td>0.54</td>
<td>0.29</td>
<td>0.20</td>
</tr>
<tr>
<td>sd(case)</td>
<td>0.16</td>
<td>0.13</td>
<td>0.20</td>
<td>0.12</td>
<td>0.11</td>
</tr>
<tr>
<td>M5: Inconsistency vs. confidence (Prediction 3): $I \sim (C - 1)$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>0.15</td>
<td>−0.20</td>
<td>0.52</td>
<td>−1.39</td>
<td>−1.72</td>
</tr>
<tr>
<td>($C - 1$)</td>
<td>−0.63</td>
<td>−0.74</td>
<td>−0.52</td>
<td>−1.66</td>
<td>−1.92</td>
</tr>
<tr>
<td>sd(expert)</td>
<td>0.54</td>
<td>0.43</td>
<td>0.67</td>
<td>0.44</td>
<td>0.27</td>
</tr>
<tr>
<td>sd(case)</td>
<td>0.62</td>
<td>0.49</td>
<td>0.81</td>
<td>0.78</td>
<td>0.61</td>
</tr>
<tr>
<td>M6: Confidence rule vs. first/second diagnoses</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>0.74</td>
<td>0.61</td>
<td>0.86</td>
<td>0.46</td>
<td>0.26</td>
</tr>
<tr>
<td>Diagnosis 1</td>
<td>−0.54</td>
<td>−0.70</td>
<td>−0.37</td>
<td>−1.11</td>
<td>−1.38</td>
</tr>
<tr>
<td>Diagnosis 2</td>
<td>−0.93</td>
<td>−1.10</td>
<td>−0.76</td>
<td>0.21</td>
<td>−0.04</td>
</tr>
<tr>
<td>sd(expert)</td>
<td>0.04</td>
<td>0.00</td>
<td>0.12</td>
<td>0.12</td>
<td>0.01</td>
</tr>
<tr>
<td>sd(case)</td>
<td>0.03</td>
<td>0.00</td>
<td>0.12</td>
<td>0.05</td>
<td>0.00</td>
</tr>
<tr>
<td>M7: Confidence rule and kind vs. wicked cases (Prediction 4)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>0.89</td>
<td>0.74</td>
<td>1.05</td>
<td>0.53</td>
<td>0.31</td>
</tr>
<tr>
<td>Wicked</td>
<td>−0.57</td>
<td>−0.89</td>
<td>−0.27</td>
<td>−0.34</td>
<td>−1.05</td>
</tr>
<tr>
<td>Diagnosis 1</td>
<td>−0.74</td>
<td>−0.93</td>
<td>−0.54</td>
<td>−1.19</td>
<td>−1.47</td>
</tr>
<tr>
<td>Diagnosis 1 × Wicked</td>
<td>0.02</td>
<td>0.19</td>
<td>1.07</td>
<td>0.57</td>
<td>−0.47</td>
</tr>
<tr>
<td>Diagnosis 2</td>
<td>−1.05</td>
<td>−1.25</td>
<td>−0.85</td>
<td>0.16</td>
<td>−0.13</td>
</tr>
<tr>
<td>Diagnosis 2 × Wicked</td>
<td>0.51</td>
<td>0.10</td>
<td>0.95</td>
<td>0.11</td>
<td>−0.92</td>
</tr>
<tr>
<td>sd(expert)</td>
<td>0.04</td>
<td>0.00</td>
<td>0.12</td>
<td>0.11</td>
<td>0.01</td>
</tr>
<tr>
<td>sd(case)</td>
<td>0.04</td>
<td>0.00</td>
<td>0.12</td>
<td>0.05</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Table S1
the basis of only those cases in which an expert’s two diagnoses for the same case differed; all other models use all data. Furthermore, model M7 only considers cases that are clearly kind ($P_i > 0.6$) or clearly wicked ($P_i < 0.4$); for the rationale for this, see the main text. The decision of the confidence rule is the reference level; that is, \textit{Diagnosis 1} and \textit{Diagnosis 2} in model M6 indicate the change in accuracy (in log odds) from the confidence rule (\textit{Intercept}) to the first or second diagnosis, respectively. In model M7, kind cases constitute the reference level; that is, \textit{Wicked} indicates for the confidence rule the change in accuracy (in log odds) when considering wicked cases instead of kind cases (the latter represented by \textit{Intercept}). Then \textit{Diagnosis 1} and \textit{Diagnosis 2} indicate for kind cases the change in accuracy (in log odds) when switching from the confidence rule (\textit{Intercept}) to the first or second diagnosis, respectively. The interaction terms (\textit{Diagnosis 1} $\times$ \textit{Wicked} and \textit{Diagnosis 2} $\times$ \textit{Wicked}) show whether the type of case (kind vs. wicked) moderates the differences between the confidence rule and first and second diagnoses, respectively.

\textbf{Results}

Regression model M2 (Table S1) shows a negative quadratic term for case ambiguity in both datasets, supporting the first prediction, that is, the higher a case’s ambiguity, the more likely an individual expert will be inconsistent when judging the same case again. Comparing the standard deviations of the group-level intercepts for experts and cases in model M1 (intercept-only model) shows that inconsistency differed more strongly for cases than for experts. Comparison of the standard deviations of the group-level intercepts for cases between model M1 (intercept-only) and M2 (incorporating a case’s ambiguity $P_i$) shows that the dispersion among cases is reduced by a factor of 5 in the mammography dataset and by a factor of 11 in the spine dataset—highlighting how much variance in inconsistency can be explained by a case’s ambiguity.

Regression model M4 (Table S1) shows a positive quadratic term for confidence in both datasets, corroborating the second prediction: the higher a case’s ambiguity (indexed by disagreement among experts’ initial diagnoses), the less confident an expert will be in their initial judgment, irrespective of whether the expert consensus for a case was correct or not. Comparing the standard deviations of the group-level intercepts, the intercept-only model M3 shows that the mean confidence of experts differs more than that of cases.

Regression model M5 (Table S1) shows a negative linear term in both datasets, corroborating the third prediction: the less confident an expert is in their initial judgment, the more likely they will be to change it when judging the same case again.

Regression model M6 (Table S1) in both datasets shows a reliably negative coefficient for \textit{Diagnosis 1}, indicating that relative to sticking with the first diagnosis, choosing the judgment with the higher confidence improves accuracy. \textit{Diagnosis 2} was only reliably negative in the mammography dataset, which indicates that the confidence rule was superior to the second diagnoses only in the mammography, but not the spine dataset—presumably because experts’ accuracy improved from the first to the second assessment phase in the latter dataset (see Figure S1 and Discussion in the main text). Model M7 (Table S1) takes into account the wickedness of cases, and partly supports the fourth prediction: For kind cases, the confidence rule was more accurate than either the first (\textit{Diagnosis 1}) or second diagnosis (\textit{Diagnosis 2}) in the mammography dataset. In the spine dataset, the confidence rule was only more accurate than the first diagnosis. The results for wicked cases were less
consistent with our fourth prediction. In both datasets using the confidence rule did not reveal a clear difference to using the first and second diagnoses (see also Figure 3 in the main text).

**Additional Results**

In the mammography dataset, the proportion of correct diagnoses was similar across first and second diagnoses (Figure S1). In the spine dataset, second diagnoses were more accurate than the first diagnoses (Figure S1).

*Figure S1.* Relation between the proportion of experts who made a correct diagnosis ($\hat{P}$) in the first vs. second diagnoses across cases. Each dot represents one case. The solid curves are LOESS smooths. Panel B employs jittering to avoid overplotting.

**References**