nature human behaviour

ARTICLES

https://doi.org/10.1038/s41562-018-0451-z

Modelling the effects of crime type and evidence on judgments about guilt

John M. Pearson[1,2,3,4], Jonathan R. Law[2,3], Jesse A. G. Skene[2,3], Donald H. Beskind[5], Neil Vidmar[5], David A. Ball[6], Artemis Malekpour[6], R. McKell Carter[7] and J. H. Pate Skene[3,4*]

Concerns over wrongful convictions have spurred an increased focus on understanding criminal justice decision-making. This study describes an experimental approach that complements conventional mock-juror experiments and case studies by providing a rapid, high-throughput screen for identifying preconceptions and biases that can influence how jurors and lawyers evaluate evidence in criminal cases. The approach combines an experimental decision task derived from marketing research with statistical modelling to explore how subjects evaluate the strength of the case against a defendant. The results show that, in the absence of explicit information about potential error rates or objective reliability, subjects tend to overweight widely used types of forensic evidence, but give much less weight than expected to a defendant's criminal history. Notably, for mock jurors, the type of crime also biases their confidence in guilt independent of the evidence. This bias is positively correlated with the seriousness of the crime. For practising prosecutors and other lawyers, the crime-type bias is much smaller, yet still correlates with the seriousness of the crime.

Growing recognition of the potential for wrongful convictions has spurred intense interest in decision-making in the criminal justice system. Retrospective studies of criminal cases have highlighted flawed forensic evidence, mistaken eyewitness identifications and defendants' prior criminal convictions as leading risk factors for error in criminal prosecutions[1–6]. In addition, some legal scholars have suggested that the risk of wrongful conviction may depend on the type of crime, with moral or emotional responses to very serious crimes increasing the risk of wrongful conviction[7–11]. These findings have triggered two approaches. The first is an extensive effort to strengthen procedures for collecting and analysing evidence, including increasing the rigour and scientific validity of common forensic methods[12–17] as well as implementing more reliable procedures for eyewitness identifications[18–21]. Second, psychologists and legal scholars have tried to better understand how jurors and others in the criminal justice system incorporate evidence and other considerations into decisions about guilt[11,14,22–38]. In this work, we extend the second approach by using a high-throughput experimental paradigm to quantify and compare the influences of evidence and potential biases on decisions by jurors and legal professionals.

Traditionally, research on criminal justice decision-making has relied on mock juror experiments using detailed case descriptions that attempt to capture as much of the context of a real trial as possible. Such studies have been widely used to investigate how jurors respond to different types of forensic evidence[22,23,25,30,33], the roles of racial and cognitive biases[39–42] and the effectiveness of jury instructions and other procedures designed to reduce those biases[43,44]. This approach has been particularly valuable for constructing psychological models of the process by which jurors incorporate multiple pieces of evidence and contextual information into a coherent judgment about guilt or innocence[36,37,45,46]. But the realistic presentation of cases in this design necessarily limits the number of independent variables that can be tested in each study.

Here, we introduce a complementary experimental approach that combines a simplified behavioural task with statistical modelling to test the effects of many evidence variables on decisions about guilt across a range of crime scenarios. In this approach, simplified descriptions of the crime and the evidence in each case make it possible to create a high-throughput behavioural task, in which each research subject evaluates many different combinations of crime-type and evidence combinations in a single session. We change a limited number of case details for each subject and then use conjoint analysis[47] from marketing to quantify attitudes in this high-dimensional crime-scenario space. This produces partially overlapping changes to the scenarios, so that we can estimate the effect size and variance for each subject, each crime scenario and each type of evidence. Data for subjects who did not see particular combinations of evidence in a particular scenario are then treated as missing and estimated using hierarchical Bayesian modelling[48,49]. We have previously applied these and similar models to analyse social behaviours in macaques[50–52] and others have shown that similar high-throughput experimental tasks using simplified scenarios can provide useful information about mechanisms underlying moral judgements in humans[53,54].

In adapting this approach to decisions to the criminal justice system, a particular concern is whether the simplified crime descriptions required by the high-throughput task design can approximate the effects of evidence and other information on decision-making in real criminal prosecutions. Simple summary descriptions of a crime and related evidence lack the context and specificity typically used in mock juror studies, to say nothing of real criminal cases. Further, in a high-throughput task, each subject evaluates a series of crime scenarios in quick succession, in contrast to the concentrated focus on a single case in a typical mock juror study or real trial. However,

[1]Department of Biostatistics and Bioinformatics, Duke University Medical Center, Durham, NC, USA. [2]Center for Cognitive Neuroscience, Duke University, Durham, NC, USA. [3]Department of Neurobiology, Duke University Medical Center, Durham, NC, USA. [4]Duke Institute for Brain Sciences, Durham, NC, USA. [5]Duke University School of Law, Durham, NC, USA. [6]Malekpour and Ball Litigation Consulting, Durham, NC, USA. [7]Department of Psychology and Neuroscience, Institute of Cognitive Science, University of Colorado, Boulder, CO, USA. *e-mail: skene@neuro.duke.edu

by averaging responses to multiple scenarios across a large number of subjects, the approach described here should, via standard statistical arguments, reliably estimate effects of interest at the population level. That is, despite its simplicity, our paradigm should be able to replicate established effects from the more realistic single-case literature and from studies of actual criminal cases while permitting a much larger number of cases and variables to be considered.

In this study, therefore, we examine how mock jurors, practising prosecutors and other practising lawyers weigh different categories of evidence and other information when deciding on guilt. Consistent with results from traditional mock juror studies and analyses of real criminal cases, we find for all groups that confidence in guilt depends most heavily on the type and amount of evidence, with physical evidence weighted more strongly than eyewitness identifications. Surprisingly, evidence of a defendant's prior criminal convictions has only a modest effect on confidence in guilt. Notably, however, we also find that the type of crime influences confidence in guilt independently of the evidence, with more serious crimes associated with increased confidence in guilt. This correlation between seriousness of the crime and confidence in guilt applies across all groups, although the magnitude of the effect is much greater for mock jurors than for prosecutors and other lawyers. The results suggest a model of decision-making in which evidence-based decisions about the guilt of a criminal defendant can be modulated by a systematic bias associated with the type of crime. More generally, the results indicate that the approach described here can complement traditional methods for investigating decision-making in the criminal justice system.

Results

Retrospective studies of real criminal cases have identified forensic evidence, eyewitness identifications and a defendant's prior criminal history as leading risk factors for wrongful convictions[1–6]. Several researchers have suggested further that the type of crime can influence judgments about a defendant's guilt[7–11,26,38,55]. In order to model the effects of crime type and evidence on judgments about guilt, we developed a high-throughput, web-based task that can sample a much larger parameter space than traditional mock juror studies[9,14,22,24–33,36,37,43,55–61]. Our task design allows each experimental subject to view 33 different crime scenarios, each of which can be coupled with a randomly selected combination of evidence, resulting in a 33 (crime type) × 3 (physical evidence) × 2 (eyewitness testimony) × 2 (prior criminal history) design (see Supplementary Information for the complete text of each option). After reading each scenario, subjects are asked to judge the strength of the case against the accused and to rate their moral or emotional responses to the crime. Figure 1a illustrates the onscreen presentation of a typical format.

Concordance with criminal justice expectations. We first administered this task to approximately 600 subjects recruited through the Amazon Mechanical Turk (mTurk) online platform[62–65]. mTurk allows subjects ('workers') to choose from a menu of online tasks, one of which was ours. To minimize differences between real jurors and our subjects, our subject pool included only adults in the United States.

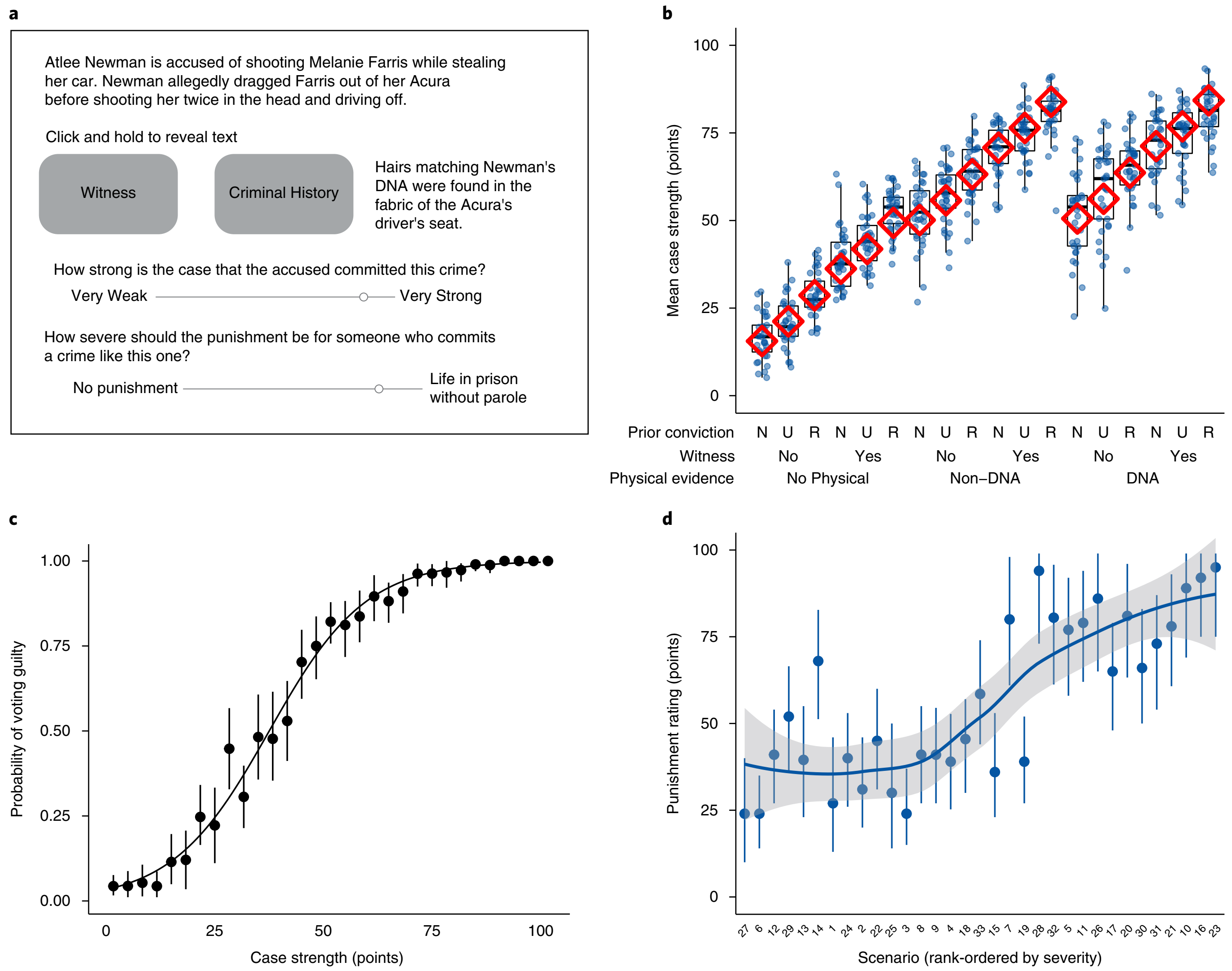
To test whether responses in our simplified experimental task resemble the expected structure of decision-making in the criminal justice system, we compared our subjects' responses to the expectations that the amount of evidence should drive confidence in guilt (Fig. 1b), stronger cases should be more likely to evoke a vote of 'guilty' (Fig. 1c) and that punishment should correspond to crime severity according to the criminal code (Fig. 1d). Our results show that the type and amount of evidence altered ratings of case strength, although the mean case strength varied for different crime scenarios (Fig. 1b). We interpret the ratings for strength of the case to reflect subjects' confidence in the guilt of the accused. To test that interpretation, we asked two subgroups of participants whether they thought the accused in each case was guilty or not guilty (Supplementary Table 1). One subgroup responded to the binary choice question in addition to rating the strength of the case, while a second group responded to the binary choice question without rating the case strength. Results were not significantly different between those groups. Combining responses across all groups, Fig. 1c makes it clear that higher ratings for case strength imply higher likelihood of responding 'guilty', suggesting that subject ratings for case strength are a reasonable proxy for confidence in guilt.

We further tested whether judgments about confidence in guilt shift over the course of the task due to viewing many crime scenarios in succession. To test this, we plotted the distribution of confidence ratings for the first five and last five scenario viewings. Because of the randomized order of presentation, each particular scenario-evidence combination appears early for some subjects and late for other subjects. We found a small shift in the distribution of responses. Confidence-in-guilt ratings were distributed more evenly for the first five viewings as compared with later ones and narrowed towards the middle of the response scale for the last five viewings (Supplementary Fig. 1). Thus, while there is some evidence that experience leads subjects to use less extreme ratings, our randomized sequence of scenario presentation shows that this effect averages out across subjects.

Finally, the 33 crime scenarios elicited a broad range of moral judgements and emotional responses. Ratings of deserved punishment, perceived threat and outrage varied with type of crime, in a rank order broadly consistent with the seriousness of the crimes under the North Carolina criminal code, which is typical of many US states (Fig. 1d and Supplementary Fig. 2).

Hierarchical and mixed model for data analysis. We used a Bayesian hierarchical model to model the effects of our experimental variables on subjects' responses[48]. This was necessary because no subject viewed every possible combination of evidence and crime: our modelling approach must account for sparsely sampled data, as well as for multiple levels of variability—both across and within individuals as well as between crimes. This approach is similar to techniques such as conjoint analysis, a method widely used in marketing research to handle situations in which the number of possible combinations of features far exceed the number of available data points per participant, hierarchical models can estimate the effects of individual features (in our case, types of evidence and type of crime) by assuming that individual responses arise from population distributions. In conjoint analysis, the emphasis is often on design: of which features to add to a product, of pricing, even of advertising. Methods for estimating the values of features have favoured multiple regression combined with adaptive sampling of cases. The key innovation of hierarchical methods, which have been part of a trend of increasing sophistication in conjoint techniques[47], is that of partial pooling: using data from the subject population to inform the fitting of a single subject's choices. For our purposes, the Bayesian hierarchical approach offers three advantages: first, Bayesian priors let us easily incorporate knowledge from previous experiments with small sample size (for example, in moving from large-scale surveys to smaller sample sizes of legal professionals). Second, Bayesian methods quantify uncertainty at multiple levels (for example, within individuals and across the population); this is crucial in studies like ours where we expect substantial individual variation. Finally, Bayesian models naturally allow us to account for missing data. In many of our datasets, respondents did not rate every case on all scales. Our Bayesian approach lets us combine such datasets into a single model with appropriately weighted uncertainties.

The primary difficulty with hierarchical models like ours is computational: the quantities of interest in most models—our beliefs



Atlee Newman is accused of shooting Melanie Farris while stealing her car. Newman allegedly dragged Farris out of her Acura before shooting her twice in the head and driving off.

Click and hold to reveal text

Hairs matching Newman's DNA were found in the fabric of the Acura's driver's seat.

How strong is the case that the accused committed this crime?

Very Weak ——○— Very Strong

How severe should the punishment be for someone who commits a crime like this one?

No punishment ——○— Life in prison without parole

a
Witness
Criminal History
b
100
75
50
25
0
Mean case strength (points)
Prior conviction N U R N U R N U R N U R N U R N U R
Witness No Yes No Yes No Yes
Physical evidence No Physical Non–DNA DNA
c
1.00
0.75
0.50
0.25
0.00
Probability of voting guilty
0
25
50
75
100
Case strength (points)
d
100
75
50
25
0
Punishment rating (points)
27 6 12 29 13 14 1 24 2 22 25 3 8 9 4 18 33 15 7 19 28 32 5 11 26 17 20 30 31 21 10 16 23
Scenario (rank-ordered by severity)

Fig. 1 | Task design and manipulation checks. a, Screenshot of the presentation for one scenario. Here the participant has clicked the box marked 'Physical evidence' to reveal one of the three alternatives for evidence in that category. **b**, Mean responses for case strength across each of the 18 possible evidence combinations. Boxplots represent variability across all 33 scenarios for fixed evidence combinations. The mean ratings for each scenario are shown as individual dots. Red diamonds illustrate mean strength estimated by our statistical model ($N=360$ subjects). The middle bar represents the median, and the lower and upper edges of the box are 25th and 75th percentiles of the data. Whiskers extend either to the range of the data in either direction or 1.5 times the interquartile range, whichever is smaller. N, no prior; U, unrelated prior; R, related prior. **c**, Case strength ratings correspond to confidence in guilt. Case strength (*x* axis) represents the mean rating across mTurk subjects. For each possible combination of scenario and evidence, the probability of voting guilty (*y* axis) reflects the percentage 'guilty' responses for a subset of subjects who were asked whether they think the accused is guilty or not guilty (Supplementary Table 7) ($N=95$ subjects). Whiskers are bootstrapped 95% confidence intervals. **d**, Ratings of deserved punishment for each crime scenario (*y* axis). Scenarios are ordered on the *x* axis according to the crime classifications under the North Carolina criminal code (see Supplementary Fig. 2 and Supplementary Table 9). Dots indicate median ratings and lower and upper whiskers are 25% and 75% of observed ratings. Grey shading represents a 95% confidence band for a local regression (LOESS smooth, blue line) of punishment as a function of scenario ($N=415$ subjects).

after having seen the data—cannot be computed efficiently from formulas. So, they are typically sampled using algorithms such as Markov chain Monte Carlo. In practice, this sampling is performed by standard analysis packages (we used the Stan Bayesian modelling language[66]) that only require users to define the model mathematically. The returned samples can then be used to compute quantities of interest such as means and variances.

In its most general form, our experiment involved multiple questions, each with multiple ratings. We modelled the data as drawn from a multivariate normal distribution with correlated ratings for, for example, punishment and confidence. We modelled the mean of each rating as a weighted sum of scenario-specific effects for each evidence type (for example, eyewitness, physical evidence). Since each rating was limited to a number between 0 and 100, with many ratings at the extreme ends of the scale, the model also accounts for the effects of censoring in our data. The model is hierarchical in that it envisions three types of variability at different levels of granularity: (1) variability across ratings within each subject (controlling for evidence), (2) variability of evidence effects across subjects for a given scenario and (3) uncertainty in evidence effects across scenarios (for example, the value of an eyewitness in arson versus robbery cases). Moreover, because we modelled these effects as Student *t*, rather than normally distributed, our inferences are less influenced by outliers. Finally, with multiple ratings, the model accounts for correlations among evidence effects. Full details of our modelling procedures are described in the Supplementary Information.

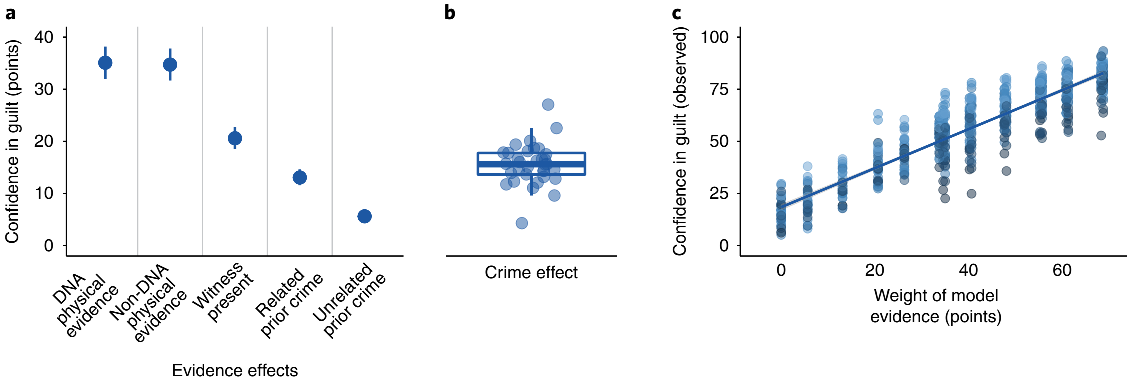


Fig. 2 | Evidence and crime effects on subject ratings for confidence in guilt. **a**, Evidence effects. Symbols represent mean effect size; error bars represent 95% credible intervals. **b**, Crime effects. Symbols represent the crime effects of individual scenarios, independent of the evidence, summarized by the box plot. Box plot parameters are the same as in Fig. 1b. **c**, Schematic illustrating increase of confidence in guilt as a function of total model evidence. Vertical groups of points represent distinct evidence combinations, with individual dots for each scenario. Dot shading indicates the variability as scenarios range from lowest crime effect (grey) to highest (light blue). As expected, the model fitting process has apportioned weights to each type of evidence such that the observed ratings are approximately linear in total evidence weight (all panels: $N = 360$ subjects).

Effects of forensic evidence, eyewitness testimony and prior convictions. Our analysis confirms that guilt confidence depends heavily on the evidence (Fig. 2). Controlling for scenario and other variables, DNA and other physical evidence linking the accused to the crime had the largest effects on subjects' guilt confidence: approximately 30 points on a 100-point scale (Fig. 2a). Contrary to scientific assessments of the relative reliability of the different types of forensic comparison methods[13,67], DNA and non-DNA physical evidence had similar effects on guilt-confidence (Fig. 2a).This suggests that our subjects viewed fibre or fingerprint evidence, for example, as nearly as dispositive as DNA. On the other hand, the presence of an identifying eyewitness had a smaller but still substantial effect (Fig. 2a). This is consistent with the extensive scientific evidence on the lower reliability of eyewitness identifications[18–21].

Compared to physical evidence or eyewitness identification, evidence of prior criminal history had only a modest effect (Fig. 2a). Our results indicate that prior conviction for a related crime increased the perceived case strength about tenpoints on a 100-point scale. A prior conviction for an unrelated crime had a smaller effect, increasing guilt confidence by about five points. While these effect sizes are relatively modest, they are statistically robust and additive with other evidence (Fig. 2a,c). These findings contrast with the widespread assumption that prior conviction evidence is powerfully prejudicial, but are consistent with statistical analyses of actual criminal cases, which found that prior-conviction evidence has a moderate but significant effect on trial outcomes[11,34,56,68]. Moreover, additional demographic information, including race, ethnicity and gender, did not affect overall ratings (Supplementary Fig. 3), suggesting that different populations used the rating scales similarly.

Despite the strong reliance on evidence, our results show further that evidence alone did not fully account for guilt-confidence (Fig. 2b). We model the overall rating for guilt-confidence as the sum of a 'crime effect' (for example, arson versus robbery) plus the increase due to each item of evidence. According to that model, in each scenario the effect of (1) the accusation against the defendant and (2) the description of the crime accounted for 6–27 points of case strength independent of the evidence (Fig. 2b). This implies that subjects in our mTurk population have, on average, a significant predisposition to believe that someone accused of a crime is guilty. The mean 'adjudicative bias'[9] for this population increases confidence in guilt by approximately 20 points, roughly equivalent to the effect of eyewitness testimony (Fig. 2b).

Different decision makers in the criminal justice system. Subjects in our mTurk study population were adults (>18 years old, mean age 37) registered with mTurk in the United States who generally lack specialized legal training (see Supplementary Tables 1–6) and thus are comparable to the jurors in the United States. Jurors, however, are not the sole decision makers in the criminal justice system. Understanding how all actors from prosecutors to defence attorneys to judges weigh evidence and integrate it with their prior beliefs and moral or emotional responses to a crime can guide efforts to optimize rules and norms in the legal system[18,19,33,53–56].

To explore potential differences between mock jurors and actual lawyers, we administered our experimental task to three groups of subjects with legal training. The first group, 52 students from three law schools, completed the study in group sessions in law school classrooms. The second and third groups, 40 practising lawyers and judges registered for a legal conference and 26 current state prosecutors, completed the task online. Figure 3 compares the results for these legally trained subjects with those of the mTurk subjects. Modelling shows that the relative ordering of effects of the different categories of evidence on confidence in guilt are strikingly similar for both, although the absolute effect sizes are somewhat higher for the legally trained subjects (Fig. 3a). Both groups gave similar weight to DNA and non-DNA physical evidence and significantly lower weight to eyewitness identifications. Prior convictions had modest effects on confidence in guilt (Fig. 3a).

In contrast, crime effects on confidence in guilt were sharply lower for the legally trained group, including prosecutors; indeed, the median crime effect across all scenarios is zero or slightly negative (Fig. 3b). While subjects could not register a negative rating for case strength, a negative crime effect represents a shift in the amount of evidence required for subjects to enter a rating greater than zero (Fig. 3c). For all three groups of legally trained subjects, the crime has a significantly lower effect on confidence in guilt than for our lay subjects. This could reflect their greater scepticism, greater reliance on the legal presumption of innocence or more analytical approach. Consistent with this, the much lower crime effect correlates with a greater effect for each type of evidence across the 33 crime scenarios (Fig. 3d). That is, for a fixed 100-point scale for confidence in guilt, the legally trained subjects allocated almost the entirety of that scale to evidence, with no indication of adjudicative bias.

Estimates of population variability are similar for all four groups, with effects most consistent across scenarios and least consistent

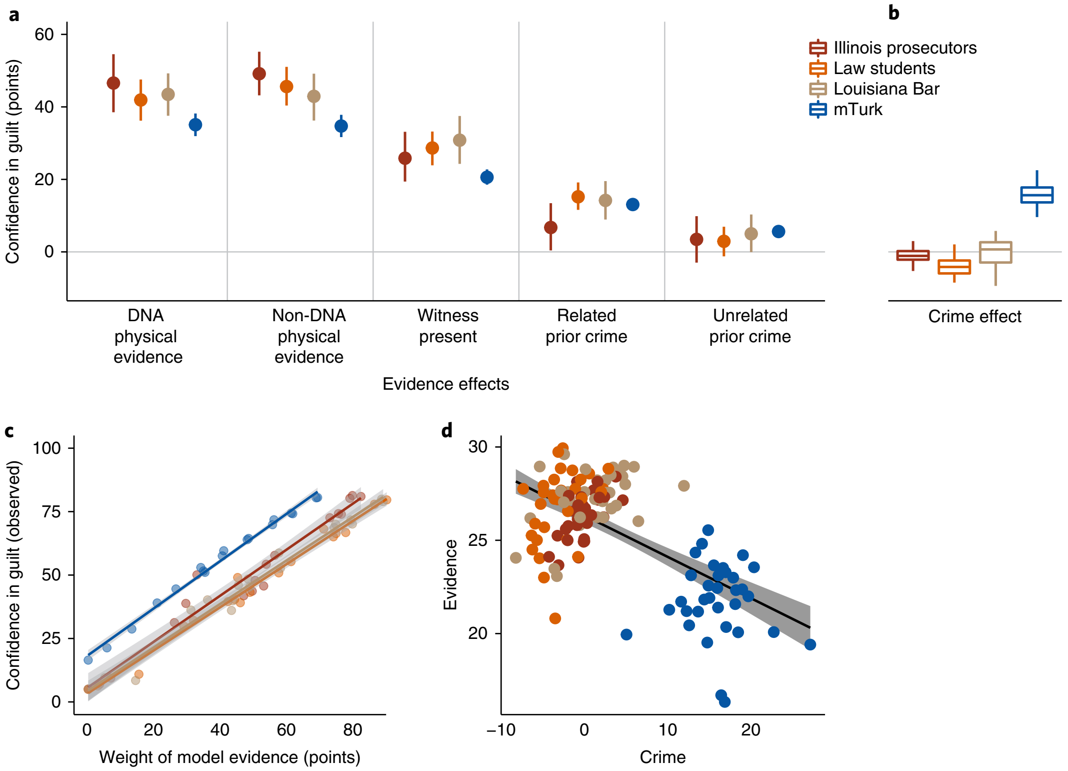


Fig. 3 | Similarities and differences between potential jurors and legally trained participants. a, The relative effect of each category of evidence on confidence in guilt is similar for potential jurors (mTurk) and three groups of legally trained participants; error bars represent 95% credible intervals. **b**, On the other hand, the crime effect (independent of evidence) is significantly smaller for the legally trained participants. **c**, As in Fig. 2c, confidence in guilt increases as a function of evidence, although potential jurors rate cases as stronger for fixed evidence than do legally trained participants. Dots indicate mean rating across cases for each evidence level. Other conventions are as in Fig. 2c. **d**, Relative contribution of evidence (*y* axis) and crime effect on confidence in guilt for each group of participants (colours as in **a**,**b**). Individual symbols represent the effect sizes for individual crime scenarios. The shaded areas in **c** and **d** represent 95% confidence bands. Given a fixed (100-point) budget, participants with legal training assigned more points to evidence and fewer to the type of crime committed (all panels: $N=26$ (Illinois prosecutors), 52 (law students), 40 (Louisiana Bar) and 360 (mTurk)).

within individuals (Supplementary Fig. 4). This suggests that ratings differences were only partly captured by differences in evidence and may have relied strongly on inferences based on individuals' perceptions of the cases.

Effects of crime type and seriousness on confidence in guilt. Our results show that the effects of both crime and evidence on confidence in guilt vary widely across the different crime scenarios (Figs. 1 and 2). By design, our scenarios differ in the type of crime described. However, they also differ in the language used to describe each case, including the names of defendants and victims, circumstances of the crime and the context for each type of evidence. Thus, the variation in confidence ratings between scenarios could represent stochastic variation in the bias associated with each crime, or the strength of the evidence in each category. Alternatively, some legal scholars have suggested that certain types of crime could be systematically associated with a higher bias or predisposition towards confidence in guilt[7,9–11,38]. This is especially important for understanding wrongful convictions. Wrongful convictions have been documented predominantly in cases involving very serious crimes such as murder or sexual assault, but it is not clear whether that is because wrongful convictions actually occur more frequently for those crimes or because errors in those cases are more likely to be detected[1,2,69].

Previous studies disagree whether a bias towards guilt for particularly serious crimes is greater than, less than, or similar to the bias in other cases[7–11,26,38,55]. Our task design allowed us to compare ratings for confidence in guilt across a much larger range of evidence and crime types than previous studies[9,26,55], including murders and sexual assaults[2,9,10,69]. To investigate whether the observed variability in crime effects could be ascribed to factors such as a crime's seriousness, we included in our model a potential correlation among regression coefficients at the population level. Figure 4a–c shows ratings on all four possible scales used by mTurk subjects, including confidence in guilt, deserved punishment, outrage and perceived threat for each case. Figure 4a shows that evidence selectively affected confidence in guilt, leaving other responses largely unaffected. Conversely, crime effects (Fig. 4b) were substantially higher for deserved punishment, outrage and perceived threat, indicating that these ratings depended predominantly on the nature of the crime itself, independent of evidence.

Figure 4c shows the correlations among these crime effects. Deserved punishment and outrage were strongly and positively correlated across cases, consistent with previous studies suggesting that the moral judgements of deserved punishment are related to emotional responses to intentional harm[53,54,70]. Notably, the crime effects on baseline confidence in guilt were positively correlated, although modestly, with deserved punishment and outrage (Fig. 4c). Credible intervals for the remaining correlations overlapped substantially with zero. The same pattern also held when our likelihood question was included (Supplementary Fig. 5).

To examine potential effects of crime seriousness on confidence in guilt, we estimated correlations among all effects (crime and evidence) for each of our populations for both confidence in guilt

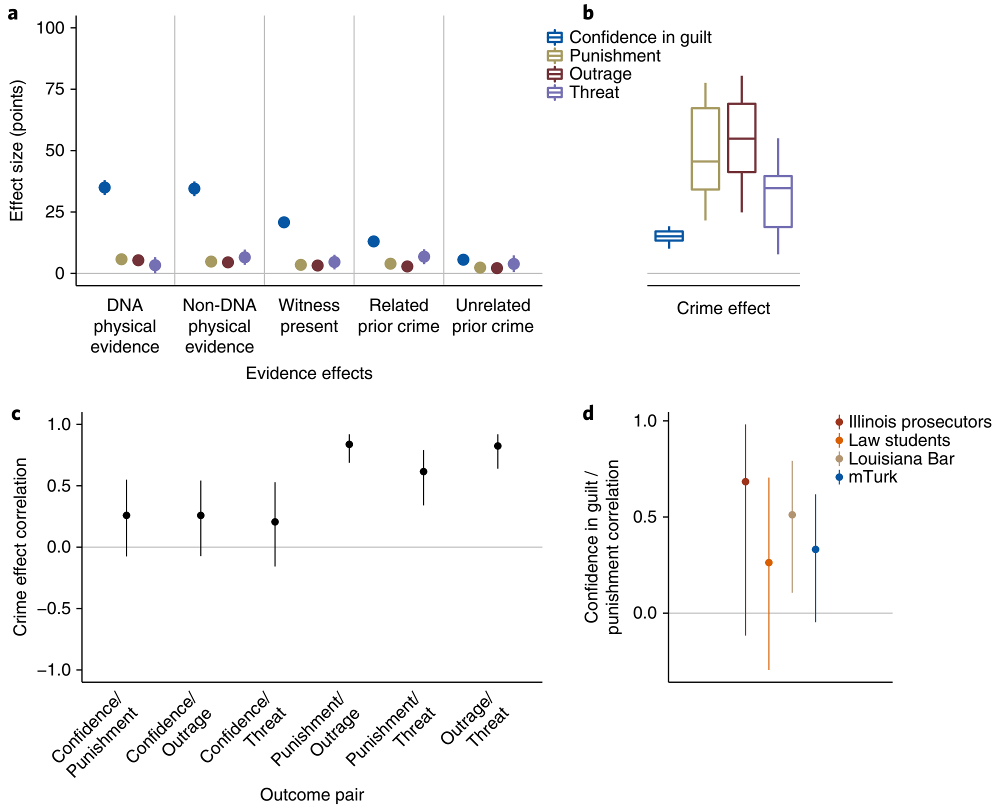


Fig. 4 | Crime effects on confidence in guilt are positively correlated with seriousness of the crime. **a**, In contrast to the effects on confidence in guilt (blue), participant ratings for deserved punishment, outrage and perceived threat are comparatively unaffected by evidence related to a particular defendant. **b**, Instead, punishment, outrage and threat ratings depend almost entirely on the crime scenarios. Box plot parameters are the same as in Fig. 1b. **c**, Crime effects on deserved punishment, outrage and perceived threat are strongly correlated with each other and positively correlated with crime effects on confidence in guilt. **d**, Crime effects for deserved punishment and case strength are positively correlated for lawyers and law students as well as mock jurors (mTurk). Error bars in **a**, **c** and **d** represent 95% credible intervals (**a–c**: *N* = 522 mTurk subjects; **d**: *N* = 26 (Illinois prosecutors), 52 (law students), 40 (Louisiana Bar) and 415 (mTurk)).

and punishment rating scales (Fig. 4d; the legally trained groups were not asked to rate outrage or threat). For each evidence effect (physical evidence, eyewitness, prior conviction), posterior model estimates of correlation were highly uncertain but mostly centred around zero (Supplementary Fig. 6). However, crime effects for case strength and punishment were modestly correlated in all groups and, for our prospective jurors, the addition of DNA evidence or an eyewitness also increased the correlation between these two measures (Supplementary Fig. 6). In other words, controlling for evidence, crimes rated as deserving greater punishment produced higher confidence in guilt. For our mock juror population, where the mean crime effect was substantial and the variability between crimes is relatively large, this correlation indicates that the 'seriousness' of a crime contributed significantly to the overall assessment of guilt. For the lawyers, while the crime effects on confidence in guilt were smaller and the mean effect is near zero, the differences between crimes remained correlated with judgments of deserved punishment, suggesting that moral intuitions may still drive variability.

Discussion

Over the last two decades, individual case studies and statistical analyses of actual criminal cases have identified some of the leading risk factors for errors in criminal prosecutions[1–5,68]. Reviews by leading scientific organizations have made important progress in addressing the reliability of forensic and eyewitness evidence[13,17,19,67]. At the same time, traditional mock-juror experiments have proven valuable for investigating how jurors evaluate that evidence and how potential biases can influence outcomes[9,10,14,22,24,27,29–31,40,42,58,68,71,72]. Here, we expand this toolbox by showing that an experimental framework widely applied in marketing and other decision research can provide numerical estimates of these effects in the context of complex decisions involving the multiple variables present in most criminal trials. This approach allows us to model complex evidence-based decisions about guilt and to identify common misperceptions or biases that are likely to have the greatest influence on decisions in the criminal justice system.

Strengths and limitations of the approach. The approach described here combines a high-throughput experimental task with hierarchical Bayesian modelling, a statistical approach well suited to sparse data and multiple levels of variability. As a complement to more conventional methods for investigating decision-making in criminal cases, our approach offers important advantages but also poses potentially significant limitations, both of which we consider here.

The advantages complement the strengths of conventional methods in criminal-justice decision research. First, the high-throughput task design makes it possible to explore the effects of multiple variables simultaneously with a large number of subjects rapidly and cost-effectively, while hierarchical modelling makes it possible to quantify these effects from sparse data. The relative speed, low-cost and flexibility of this approach enables multiple iterations of an experiment to optimize the task or to test the effects of minor variations in case presentation, questions and presentation format (see Supplementary Information). These advantages suggest that our approach can complement more conventional approaches, providing a relatively rapid and cost-effective first pass that (1) screens variables and their potential interactions, (2) tests alternative hypotheses and (3) prioritizes issues for more time- and resource-intensive studies.

The second advantage of our approach is that it places these effects within a conceptual and computational framework that applies to a wide range of decisions in humans and other animals[48–54]. Because our experimental task design and computational models are compatible with functional brain imaging, electroencephalography and other functional measures, it offers a critical link between more realistic methods that establish the ecological validity of a critical influence on decision-making as well as on the ability to investigate brain mechanism that mediate those effects.

The third advantage of our approach, its ability to obtain robust measures for multiple variables from sparse data sets, may prove especially valuable for investigating decision-making by active prosecutors and other key decision makers in the criminal justice system who rarely have been represented in studies that rely on more time-intensive experimental designs.

Despite these potential benefits, the requirements of a high-throughput task design impose significant limitations that could compromise its ecological validity, limiting our ability to draw conclusions about real-world decisions. For example, our approach requires each subject to view a relatively large number of individual crime scenarios in succession, with each scenario consisting of short descriptions of the crime and each item of evidence with minimal detail. By contrast, jurors in a criminal trial focus exclusively on one case, with evidence covered in rich detail. Traditional mock juror experiments also typically focus on a single case, in which the crime and evidence are presented as detailed as possible to what jurors will hear in trial. Finally, for this study, we recruited mock jurors through the mTurk platform. As others have found[56,62–65,73], the mTurk population cannot be considered representative of actual jurors and we cannot control the test environment for mTurk subjects participating in our study online.

We investigated each of these potential limitations. With regard to the validity of results obtained with mTurk subjects, we compared results for the mTurk subjects with results for law students who took part in our study in live events in their law school classrooms, with actual prosecutors and with a separate cohort of practising lawyers. We found that the relative effects of forensic evidence, eyewitness identifications and prior-conviction evidence were robust across individual 'batches' of mTurk subjects who viewed different variations of the behavioural task and between mTurk subjects and our law students, prosecutors and lawyers (Fig. 3a). Similarly, both the lawyers and mTurk subjects showed a positive correlation between confidence in guilt and the seriousness of a crime (Fig. 4d), although the magnitude was much smaller in the legally trained populations (Fig. 3b). As others have cautioned for mock-juror studies using undergraduate students or mTurk subjects[9,56,74,75], the differences we observe between mock jurors recruited through mTurk and real law students and lawyers need to be confirmed with further studies of actual jurors. On the other hand, the striking similarities across groups for the relative effects of forensic evidence, eyewitness identifications, prior conviction evidence and crime seriousness indicate that the majority of effects described here are applicable to a broad population, including real participants in the criminal justice system.

We also examined possible effects of our task design in which each subject evaluates multiple cases in succession (Supplementary Fig. 1). The results show that subjects do tend to adjust their rating scale over the course of the task, but we are able to control for this by randomizing the order in which scenarios are presented to individual subjects and averaging the main effects across subjects.

Together, these results show that our experimental design can produce statistically robust estimates for the effects of evidence and crime type on judgments about guilt across multiple crime scenarios and subject populations. A more fundamental question, however, is whether the results adequately reflect/capture (reproduce) the effects of these factors on real-world criminal justice decisions. Here, we have examined this by modelling the effects of variables previously identified as risk factors for wrongful conviction, then comparing our results with results from conventional mock juror experiments and studies of actual criminal cases. The results are generally consistent with the earlier findings, while expanding the results across a broader range of crime types and subjects.

Forensic evidence and eyewitnesses. We first examined the effects of evidence based on forensic methods long regarded as highly reliable, but whose reliability has been challenged recently by scientific evaluations[12,13,17,67]. Reviews by leading scientific organizations over the last decade have concluded that many widely used forensic methods either lack adequate scientific foundations or have a greater potential for error than has been commonly recognized[12,13,17,67,76]. In particular, these reviews have distinguished the extensive empirical validation of conventional DNA identification methods compared to other feature-comparison methods, including latent fingerprint analysis[13,67]. Our results indicate that, absent specific testimony about different forensic science methods and the accuracy or uncertainty of the conclusion in a particular case, our mock jurors give much greater weight to forensic science evidence than to other types of evidence, but make little distinction between the reliability of DNA identification and non-DNA comparisons (Fig. 2). Law students, practising lawyers and prosecutors in our samples did the same (Fig. 3a). Because we presented evidence in the form of short, conclusory statements, our study subjects had to rely on their prior beliefs about the reliability of each forensic method. Our results indicate that these beliefs correspond more closely with longstanding beliefs about the reliability of traditional forensic sciences than with more recent scientific studies. Jurors' confidence in the reliability of forensic methods may have been reinforced by popular media, as in the 'CSI effect'[23,25,30]. Consistent with our findings, previous studies have found that jurors tend to give excessive weight to conclusions based on traditional forensic science methods unless the jurors receive explicit information about uncertainty or the potential for error for each method[14,22,29,33].

Like many forensic methods, eyewitness testimony was long regarded as highly reliable in criminal prosecutions. But our results indicate that, for all subjects, eyewitness identifications have a much smaller effect on confidence in guilt than does physical evidence (Fig. 3a). The reasons are unclear. One possibility is the much longer history of scientific research on eyewitness testimony. Eyewitness identification has been the subject of more than 30 years of experimental studies testing reliability, including factors that affect the accuracy of encoding, recall and identification[18–21]. The results have been widely disseminated in both the popular press and the criminal justice system, resulting in altered procedures for collecting identification but also more effective defence challenges[21,58,72]. Our results suggest that this limited reliability has begun to penetrate popular awareness, affecting the prior beliefs or assumptions about eyewitness evidence. Consistent with these results, surveys have found that jurors, judges and the general public recognize many factors that can reduce eyewitness reliability[58,72], although both lay subjects and lawyers erred in identifying how those factors affect accuracy.

Prior criminal history. One of striking contrast between our results and conventional legal wisdom is that evidence of a previous conviction, even for a related crime, has only a modest effect on confidence in guilt (Fig. 2a). Prior conviction evidence is widely regarded by courts and practising attorneys as powerfully prejudicial against a defendant[3,5]. Rules of evidence, for example, generally exclude evidence of prior convictions, with limited exceptions. One key exception allows prior convictions evidence to be admitted if the defendant testifies. As a results, concern over the prejudicial effects of a prior conviction can affect strategic decisions about whether a defendant with a criminal record should testify or even whether he should plead guilty[3,34,68].

Our results, however, indicate that evidence of a prior conviction for a related crime increases overall confidence in guilt by only about ten points on a 100-point scale. While this contrasts with the conventional wisdom, it is consistent with studies of actual criminal cases, which have found a similarly limited effect of prior conviction evidence on the outcomes in real trials[11,34,38,56,68]. Despite the modest magnitude of the effect, our results show that the effect of a prior criminal conviction is statistically robust and additive with other evidence (Fig. 2a,c). This suggests that introducing evidence of prior conviction can have a significant effect on the likelihood of conviction in some cases, most notably where other evidence is moderately strong but not compelling. As illustrated in Fig. 1c, when the overall strength of the case is very weak or very strong, a ten-point increase in confidence will not produce a substantial increase in the percentage of mock jurors who would find the defendant guilty. On the other hand, where the other evidence in ambiguous, a ten-point increase in confidence can have a much larger effect on the likelihood of voting guilt (Fig. 1c). This, too, is consistent with previous statistical analyses of real criminal cases, which found that prior conviction evidence has little effect when the other evidence is very strong or very weak, but can tilt the balance towards conviction when other evidence is ambiguous[11,34,38,56]. Thus, while our result contradicts conventional wisdom among experienced lawyers, it closely matches the results from empirical studies of actual criminal trials.

At the same time, our results shed light on previous findings that prior conviction evidence only affects trial outcomes when other evidence is ambiguous. Kalven and Zeisel's classic 'liberation hypothesis', for example, proposed that ambiguous factual evidence frees jurors to consider other information about a defendant, including prior convictions[38]. Yet in our study, prior convictions simply add to the total of other evidence (Fig. 1b). These findings can be reconciled by observing that for cases in which the evidence is not clearly dispositive, small effects such as the existence of a prior conviction can be decisive with regard to finding a defendant guilty (Fig. 1c). Our results do not exclude an additional interaction that depends on the strength of other evidence, but they do suggest that the prior conviction evidence has a main effect independent of the other evidence.

Biases independent of evidence. In addition to the effects of evidence, our results show that confidence in guilt can be influenced by other factors. In particular, for our mTurk respondents, an accusation and description of the crime—without any evidence—increased confidence in guilt by 13–25 points. This effect was positive for all crime scenarios, suggesting an overall bias towards guilt[9,23]. Each piece of evidence then increased confidence in guilt. By contrast, the average bias was much lower for our legally trained subjects. Notably, this difference does not appear to reflect a defence-favouring bias for the legally trained groups, although it reflected a prosecution-favouring bias for the mock jurors[9,23]. The lower bias for lawyers and law students may reflect legal training and case-analysis experience, rather than a generic bias towards either prosecution or defence.

For all groups, crime effects on confidence in guilt varied significantly among the individual scenarios. This variation was largest for the mock jurors (12 points), lower for practising lawyers and law students (6–7 points) and lowest for active prosecutors (2 points). Previous studies disagree as to whether the perceived seriousness of a crime increases or decreases baseline confidence in guilt[7–11,26,55]. Our task design allowed us to compare subjects' confidence in guilt across a broad range of evidence and crime types, including murders, sexual assaults[9,10] and a wide range of other felonies and misdemeanours. For the mock jurors and lawyers, crime effects were positively correlated with the seriousness of the crime, as reflected in their ratings for deserved punishment. Thus, baseline confidence in guilt was significantly lower than the average for relatively minor crimes, while crimes rated as deserving the most punishment also elicited the highest effects on confidence in guilt. For the most serious crimes, including murders and sexual assaults, the effect of the accusation and crime description was comparable to the effect of an eyewitness identification. The smaller effect sizes and small sample sizes for our legally trained subjects do not allow us to conclude whether this differs across groups, although for each group the estimate of this correlation is positive and excludes zero. Thus, the results for all groups are inconsistent with a negative effect of crime seriousness on confidence in guilt[26,55], but consistent with statistical studies of actual cases, which have found a higher likelihood of conviction for defendants accused of more serious crimes[11].

Our results suggest that mock jurors and the lawyers in our study share an underlying bias towards greater confidence in guilt for more serious crimes. This has the potential to affect decisions throughout a criminal investigation and prosecution. At least in the context of our simplified experimental task, the effect of this crime-type bias is much greater for mock jurors than for prosecutors and other lawyers. The reasons for this difference are not yet clear. But evidence for this underlying bias suggests that it would be valuable to investigate conditions and interventions that can mitigate or exacerbate the effect of this bias in the context of a real criminal investigation and trial.

Conclusion

Combining a high-throughput, flexible task design with statistical modelling designed for sparse data sets, the approach described here offers a potentially valuable complement to more traditional methods for investigating decisions in the criminal justice system. Clearly, one must exercise caution in translating results from this kind of study into practice. But the similar effects of different evidence types and of crime type on judgments of guilt by mTurk participants, law students and experienced lawyers suggest that high-throughput studies using the mTurk population can be effective in capturing features of legal decision-making that are robust across different groups and test conditions. At the same time, the notable difference in effect size for the crime type bias between mTurk participants and legally trained subjects illustrates the ability of this approach to identify effects of training, experience and other factors on decision-making by different actors in the criminal justice system. Our study design's ability to capture important risk factors and elements of decision-making as identified in retrospective studies of actual cases[3,4,34,68] indicates that this type of high-throughput experimental design can provide a rapid initial screen for testing new hypotheses, identifying potential interactions and assessing proposed procedural reforms and best practices. Results from this type of rapid screen can prioritize hypotheses to be tested in more time- and resource-intensive studies, including focus groups, traditional experiments with mock or actual jurors and retrospective case studies.

At the same time, our high-throughput task format and computational modelling is compatible with models and methods that have been widely used for investigating other kinds of evidence-based decisions, including complex, multi-attribute choices in marketing research[47,48] and public policy proposals[49]. Our results show that these decision models can be useful for characterizing complex decision-making by different actors in the criminal justice system. Further, because the task design and modelling are compatible with functional brain imaging and electrophysiology, the approach described here offers a means for investigating the neural mechanisms that mediate effects identified as significant risk factors for errors in more realistic mock juror experiments and in real criminal cases.

Methods

Subjects, recruitment and sampling. These studies were approved by the Duke University Institutional Review Board. All participants provided informed consent and the study complied with all relevant ethical regulations. The behavioural task described in the text was administered via a web app implemented in node.js (code available from the authors on request). Subjects were recruited through Amazon mTurk, by posted notices in law schools and by direct email to members of cooperating prosecutors' offices and attendees at legal conferences. Subjects recruited through mTurk were adults based in the United States, according to mTurk records and IP addresses. All subjects accessed the task online through the web app site. At the conclusion of the task, subjects were routed to a separate site to complete a demographic questionnaire.

Human participants were all healthy adults. Participants recruited online through mTurk indicated that they were at least 18 years old and registered with mTurk from US addresses. A demographic breakdown is included in Supplementary Tables 1–6. Law students were second- and third-year students recruited from three US law schools (Duke, Wake Forest and University of Colorado-Boulder). Practising prosecutors were members of the Lake County (Illinois) State's Attorney's office. Other practising attorneys were recruited through email requests to participants in a continuing legal education conference sponsored by the Louisiana Bar Association. For each of these legally trained participant groups, completion of the study was purely opt-in. After all exclusions (described below), our full sample included 759 mTurk participants, 52 law students, 26 prosecutors and 40 other lawyers (Supplementary Table 7).

Data were collected via Qualtrics (for demographic data) and a custom web application designed by the research team. Participants viewed the study within a web browser. Data collection and analysis were not performed blind to the conditions of the experiment. However, members of the research team were only indirectly involved in data collection: for mTurk participants, all data were collected remotely without the involvement of the research team. For law students, a member of the research team was present to introduce the study but left the room prior to data collection.

Sample sizes for mTurk studies were solicited in approximately 50 participant blocks for variations on the basic study design. By using Bayesian methods, we are able to combine across samples to calculate a final uncertainty about the estimands of interest. Because we are not performing sequential hypothesis testing and because we use all non-excluded data together in estimating effects, this multi-stage sampling does not induce bias in our conclusions[77].

Experimental task design. The task for this study uses a 33 × 3 × 2 × 3 experimental design. Subjects view 33 scenarios that each describe accusation of a specific crime. The 33 crimes span shoplifting to rape, murder and child sexual abuse. Each subject views the presentation of each crime type (a 'scenario') in a text box. It remains onscreen until the subject has answered a series of questions and then clicks a button to move to the next scenario. This initial description contains no evidence of who committed the crime. While type of crime is the primary difference between scenarios, the scenarios also differ in other details, including names of defendants and victims, and the circumstances of the crime. Varying these other details is designed to keep subjects engaged throughout the task and to encourage subjects to treat each scenario as a distinct crime. At the same time, the varied descriptions raise the potential for differences other than crime type to influence effects of the crime scenario on judgments of guilt. However, our large number of scenarios and the range of crime seriousness across those scenarios allows us to test specifically for a main effect of crime seriousness on judgments of guilt. The text for each scenario is listed in the Supplementary Information.

Below the descriptive paragraph for each scenario, subjects see one or more categories of evidence implicating the named suspect (Fig. 1a). In our standard format, subjects chose which evidence to view first, second and third clicking on one of three boxes labelled 'Physical evidence', 'Witness' and 'Criminal history'. Clicking on each box shows one of the following: (1) one of three physical evidence possibilities that link the accused to the crime (no physical evidence, DNA evidence or non-DNA physical evidence such as fingerprint or ballistic evidence); (2) either of two kinds of 'witness' (eyewitness or a non-eyewitness) or (3) one of three options of criminal history (no prior convictions, prior conviction for a related crime or prior conviction for an unrelated crime). This results in 18 unique evidence combinations (3 × 2 × 3) for each crime scenario. Each subject sees all 33 crime scenarios paired with only one randomized combination of evidence. Although each subject sees only one randomized combination of evidence for each crime scenario, over our large number of subjects we can test each of the 33 crime scenarios with all 18 evidence combinations (594 unique combinations).

Following the scenario and evidence boxes, subjects are asked to judge the strength of the case against the accused and to rate their moral or emotional responses to the crime.

By design, the presentation of evidence in this study differs in several respects from the way evidence is typically presented to jurors in an actual trial or in traditional mock juries. First, our format lets subjects choose the sequence order in which to view the evidence in each scenario. By contrast, jurors at trial typically hear the evidence in the sequence determined by the prosecution, and current models of juror decision-making such as the Story Model emphasize the influence of the order of presentation on the weight jurors assign to individual items of evidence[36,37,45,46,78]. In addition, each item of evidence in this study consists of a short, conclusory statement with no details of how the evidence was collected or analysed, thus requiring subjects to use their own prior beliefs and assumptions to decide how much weight to give that evidence.

Within each evidence category, furthermore, the specific context and description of the evidence in each category varies for the individual scenarios (Supplementary Information); thus, the objective strength of the evidence in a particular category can be stronger for some scenarios than others. In the majority of our scenarios, for example, the source of the DNA evidence is blood, skin, saliva or semen, all of which typically allow comparison of nuclear DNA. In six scenarios, however, the source of the DNA evidence is described as hair samples only (Supplementary Information). Because hair samples typically permit analysis of mitochondrial DNA only, the objective strength of the DNA evidence is weaker for those scenarios. Similarly, the non-DNA physical evidence in the majority of scenarios consists of a fingerprint match, while the non-DNA physical evidence in five scenarios consists of hair, fibre or paint comparisons. Among the widely used non-DNA feature-comparison methods, latent fingerprint evidence is traditionally regarded as the most reliable and has been most rigorously tested for scientific validity than other feature-comparison methods[13,19,67,79]. This can potentially increase variation between scenarios in the effect of each evidence category on judgments of guilt. At the same time, the empirical studies show that fingerprint analysis is subject to a higher error rate and greater sources of variation than conventional DNA methods[13,19,67,79]. Thus, averaging across scenarios, the mean strength of the DNA evidence is objectively much greater than the corresponding non-DNA physical evidence.

While these features of the current study limit the realism of our experimental task, they make it possible to identify main effects of each evidence category at a population level, independent of variation in the sequence of presentation and the specific context of the evidence in each particular case. Further, our task format easily accommodates future iterations of this or similar studies to investigate how specific changes in the case presentations alter these main effects. It is important to remember that while our paradigm may yield less accurate predictions in individual cases, by increasing drastically the numbers of subjects and cases we collect, we are leveraging additional statistical power to derive much more accurate inferences in aggregate.

Study variants. Across a series of studies, we varied the task format and the number of questions posed to different groups of subjects (Supplementary Table 7). These variations showed no significant effect on our subjects' responses in cases where evidence was presented, so we combined data across experiments (see below).

Randomization and exclusion criteria. All participants responded to all 33 scenarios, but the particular evidence combination for each scenario was determined by a random permutation of all 18 possibilities. This randomization was performed once per subject at the beginning of each experiment using a Knuth shuffle method.

We excluded participants whose patterns of response either indicated a consistent failure to move the rating slider or unusually low time to complete the survey. Specifically, all responses from a subject were excluded if the data set met one of the following criteria:

(1) Subject did not complete the entire task
(2) Subject did not complete all 33 questions in test portion
(3) Subject completed the test portion but not the demographics portion
(4) Data suggested that the subject did not participate in the task, or stopped participating in the task, due to repeated lack of change to default response values
(5) Unidentified and un-remediated coding errors that resulted in some subjects who were presented with the same question more than four times
(6) Subject had participated in a previous version of the experiment (verified via mTurk user name).

Partial (repeat) data for an individual subject were excluded if the data set met the following criteria: unidentified and un-remediated coding error that resulted in some subjects being presented with same question two, three or four times. In those cases, responses after the first response were redacted. Exclusion criteria were not established in the initial study design. We established the exclusion criteria after reviewing responses to the first round of data collection, each group comprising 50–100 subjects (Supplementary Table 7). Those criteria were then applied to all subsequent rounds of data collection.

Missing data and combining across experiments. Because we ran multiple versions of the experiment with our mTurk population, many subjects were only required to supply ratings for one or two of our six outcome types. Nevertheless, in modelling these data, we proceeded as if each subject possessed a full set of outcomes, but these outcomes were only partially observed. This is accounted for by the Bayesian methods we employ[80], which do not require that we either impute missing data or restrict ourselves to complete cases. Thus, we were able to include in our multivariate outcome model all data points with *any* of the five continuous ratings and these were all used to inform inferences about model parameters. Where we have only small overlap between some pairs of observed outcomes (for example, confidence and threat) the resulting posterior uncertainties are inevitably larger but are nonetheless correct summaries of our beliefs given the observed data.

For this reason, for many of our analyses, we were able to combine data across models. For example, in considering correlations among outcome types, we did not restrict ourselves to the single experiment in which subjects provided all six response types. Rather, we included all data in which subjects provided *any* of the five continuous ratings. A full description of our data inclusion criteria for each analysis performed is given in Supplementary Table 8.

Model fitting. We modelled the data as described in the text, fitting the models using Hamiltonian Markov chain Monte Carlo as implemented in the Stan Bayesian modelling language[66]. Full details of the models and inference procedure are available in Supplementary Information.

Reporting Summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this article.

Code availability. All code to reproduce the results of this paper, including models, figures, supplementary figures and supplementary tables, is available at https://github.com/pearsonlab/legal.

Data availability

All data used in the analyses are available and documented at https://github.com/pearsonlab/legal.

Received: 22 January 2018; Accepted: 12 September 2018;
Published online: 29 October 2018

References

1. Gross, S. R. & O'Brien, B. Frequency and predictors of false conviction: why we know so little, and new data on capital cases. *J. Empir. Leg. Stud.* **5**, 927–962 (2008).
2. Gross, S. R. Convicting the innocent. *Annu. Rev. Law Soc. Sci.* **4**, 173–192 (2008).
3. Gould, J. B., Carrano, J., Leo, R. A. & Hail-Jares, K. Predicting erroneous convictions. *Iowa Law Rev.* **99**, 471–522 (2014).
4. Garrett, B. L. & Neufeld, P. J. Invalid forensic science testimony and wrongful convictions. *Va. Law Rev.* **95**, 1–97 (2009).
5. Garrett, B. L. *Convicting the Innocent. Where Criminal Prosecutions Go Wrong* (Harvard Univ. Press, Cambridge, 2011).
6. Cole, S. A. Forensic science and wrongful convictions: from exposer to contributor to corrector. *N. Engl. Law Rev.* **46**, 711–736 (2011).
7. Vidmar, N. Case studies of pre- and midtrial prejudice in criminal and civil litigation. *Law Hum. Behav.* **26**, 73–105 (2002).
8. Vidmar, N. When all of us are victims: juror prejudice and terrorist trials. *Chic. Kent Law Rev.* **78**, 1143–1178 (2003).
9. Wiener, R. L., Arnot, L., Winter, R. & Redmond, B. Generic prejudice in the law: sexual assault and homicide. *Basic Appl. Soc. Psych.* **28**, 145–155 (2006).
10. Vidmar, N. Generic prejudice and the presumption of guilt in sex abuse trials. *Law Hum. Behav.* **21**, 5–25 (1997).
11. Gastwirth, J. L. & Sinclair, M. D. A re-examination of the 1966 Kalven-Zeisel study of judge-jury agreements and disagreements and their causes. *Law Probab. Risk* **3**, 169–191 (2004).
12. Fabricant, M. C. & Carrington, T. The shifted paradigm: forensic science's overdue evolution from magic to law. *Va J. Crim. Law* **4**, 1–115 (2016).
13. Lander, E. S. et al. *Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods* (Executive Office of The President's Council of Advisors on Science and Technology, 2016); https://obamawhitehouse.archives.gov/administration/eop/ostp/pcast/docsreports
14. McQuiston-Surrett, D. & Saks, M. J. The testimony of forensic identification science: what expert witnesses say and what factfinders hear. *Law Hum. Behav.* **33**, 436–453 (2009).
15. Morrison, G. S. Special issue on measuring and reporting the precision of forensic likelihood ratios: introduction to the debate. *Sci. Justice* **56**, 371–373 (2016).
16. Ulery, B. T., Hicklin, R. A., Buscaglia, J. & Roberts, M. A. Accuracy and reliability of forensic latent fingerprint decisions. *Proc. Natl Acad. Sci. USA* **108**, 7733–7738 (2011).
17. National Commission on Forensic Sciences *Reflecting Back—Looking Toward the Future* (US Department of Justice, 2017); https://www.justice.gov/archives/ncfs/page/file/959356/download
18. Albright, T. D. Why eyewitnesses fail. *Proc. Natl Acad. Sci. USA* **114**, 7758–7764 (2017).
19. National Reseach Council. *Identifying the Culprit: Assessing Eyewitness Identification* (National Academies Press, 2014); https://doi.org/10.17226/18891
20. Steblay, N. K. Scientific advances in eyewitness identification evidence. *William Mitchell Law Rev.* **41**, 1090–1127 (2015).
21. Wells, G. L. in *Modern Scientific Evidence: The Law and Science of Expert Testimony* Vol. 2 (eds Faigman, D. L. et al.) 615–662 (Thomson/West, Eagan, 2016).
22. Thompson, W. C., Kaasa, S. O. & Peterson, T. Do jurors give appropriate weight to forensic identification evidence? *J. Empir. Leg. Stud.* **10**, 359–397 (2013).
23. Smith, L. L. & Bull, R. Identifying and measuring juror pre-trial bias for forensic evidence: development and validation of the Forensic Evidence Evaluation Bias Scale. *Psychol. Crime Law* **18**, 797–815 (2012).
24. Smith, L. L., Bull, R. & Holliday, R. Understanding juror perceptions of forensic evidence: investigating the impact of case context on perceptions of forensic evidence strength. *J. Forensic. Sci.* **56**, 409–414 (2011).
25. Schweitzer, N. J. & Saks, M. J. The CSI effect: popular fiction about forensic science affects the public's expectations about real forensic science. *Jurimetrics* **47**, 357–364 (2007).
26. Rind, B., Jaeger, M. & Strohmetz, D. B. Effect of crime seriousness on simulated jurors' use of inadmissible evidence. *J. Soc. Psychol.* **135**, 417–424 (1995).
27. Nance, D. A. & Morris, S. B. Juror understanding of DNA evidence: an empirical assessment of presentation formats for trace evidence with a relatively small random-match probability. *J. Legal. Stud.* **34**, 395–444 (2005).
28. Martire, K. A., Kemp, R. I., Sayle, M. & Newell, B. R. On the interpretation of likelihood ratios in forensic science evidence: presentation formats and the weak evidence effect. *Forensic. Sci. Int.* **240**, 61–68 (2014).
29. Lieberman, J. D., Carrell, C. A., Miethe, T. D. & Krauss, D. A. Gold versus platinum: do jurors recognize the superiority and limitations of DNA evidence compared to other types of forensic evidence? *Psychol. Public. Policy Law.* **14**, 27–62 (2008).
30. Kim, Y. S., Barak, G. & Shelton, D. E. Examining the 'CSI-effect' in the cases of circumstantial evidence and eyewitness testimony: multivariate and path analyses. *J. Crim. Justice.* **37**, 452–460 (2009).
31. Kaye, D. H., Hans, V. P., Dann, B. M., Farley, E. & Albertson, S. Statistics in the jury box: how jurors respond to mitochondrial DNA match probabilities. *J. Empir. Leg. Stud.* **4**, 797–834 (2007).
32. Kassin, S. M., Dror, I. E. & Kukucka, J. The forensic confirmation bias: problems, perspectives, and proposed solutions. *J. Appl. Res. Mem. Cogn.* **2**, 42–52 (2013).
33. Garrett, B. & Mitchell, G. How jurors evaluate fingerprint evidence: the relative importance of match language, method information, and error acknowledgment. *J. Empir. Leg. Stud.* **10**, 484–511 (2013).
34. Eisenberg, T. & Hans, V. P. Taking a stand on taking the stand: the effect of a prior criminal record on the decision to testify and on trial outcomes. *Cornell Law. Rev.* **94**, 1353–1390 (2009).
35. Danziger, S., Levav, J. & Avnaim-Pesso, L. Extraneous factors in judicial decisions. *Proc. Natl Acad. Sci. USA* **108**, 6889–6892 (2011).
36. Hastie, R. & Pennington, N. *Psychology of Learning and Motivation* Vol. 32 (eds Busemeyer, J., Hastie, R. & Medin, D. L.) 1–31 (Academic Press, San Diego, 1995).
37. Pennington, N. & Hastie, R. Explaining the evidence: tests of the Story Model for juror decision making. *J. Pers. Soc. Psychol.* **62**, 189–206 (1992).
38. Kalven, H. & Zeisel, H. *The American Jury* (Little, Brown & Company, Chicago, 1966).
39. Young, D. M., Levinson, J. D. & Sinnett, S. Innocent until primed: mock jurors' racially biased response to the presumption of innocence. *PLoS ONE* **9**, e92365 (2014).
40. Simon, D. *In Doubt: The Psychology of the Criminal Justice Process* (Harvard Univ. Press, Cambridge, 2012).
41. O'Brien, B. & Findley, K. *Psychological Perspectives: Cognition and Decision Making. Examining Wrongful Convictions: Stepping Back, Moving Forward* (Carolina Academic Press, Durham, 2014).
42. O'Brien, B. Prime suspect: an examination of factors that aggravate and counteract confirmation bias in criminal investigations. *Psychol. Public. Policy Law.* **15**, 315–334 (2009).
43. Steblay, N., Hosch, H. M., Culhane, S. E. & McWethy, A. The impact on juror verdicts of judicial instruction to disregard inadmissible evidence: a meta-analysis. *Law Hum. Behav.* **30**, 469–492 (2006).
44. O'Brien, B. A recipe for bias: an empirical look at the interplay between institutional incentives and bounded rationality in prosecutorial decision making. *Miss. Law Rev.* **74**, 999–1050 (2009).
45. Hastie, R. in *Better Than Conscious? Decision Making, the Human Mind, and Implications for Institutions* (eds Engel, C. & Singer, W.) 371–390 (MIT Press, Cambridge, 2008).

46. Pennington, N. & Hastie, R. Evidence evaluation in complex decision making. *J. Pers. Soc. Psychol.* **51**, 242–258 (1986).
47. Gustafsson, A., Harrman, A. & Huber, F. *Conjoint Measurement: Methods and Applications* (Springer, Berlin, Heidelberg, New York, 2013).
48. Gelman, A. & Hill, J. *Data Analysis Using Regression and Multilevel/Hierarchical Models* (Cambridge Univ. Press, New York, 2006).
49. Gelman, A., Fagan, J. & Kiss, A. An analysis of the New York City police department's 'stop-and-frisk' policy in the context of claims of racial bias. *J. Am. Stat. Assoc.* **102**, 813–823 (2007).
50. Brent, L. J. N. et al. Genetic origins of social networks in rhesus macaques. *Sci. Rep.* **3**, 1042 (2013).
51. Chang, S. W. C. et al. Neural mechanisms of social decision-making in the primate amygdala. *Proc. Natl Acad. Sci. USA* **112**, 16012–16017 (2015).
52. Watson, K. K. et al. Genetic influences on social attention in free-ranging rhesus macaques. *Anim. Behav.* **103**, 267–275 (2015).
53. Buckholtz, J. W. & Marois, R. The roots of modern justice: cognitive and neural foundations of social norms and their enforcement. *Nat. Neurosci.* **15**, 655–661 (2012).
54. Treadway, M. T. et al. Corticolimbic gating of emotion-driven punishment. *Nat. Neurosci.* **17**, 1270–1275 (2014).
55. Kerr, N. L. Severity of prescribed penalty and mock jurors' verdicts. *J. Pers. Soc. Psychol.* **36**, 1431–1442 (1978).
56. Bellin, J. The silence penalty. *Iowa Law Rev.* **103**, 395–434 (2018).
57. Jones, A. M. & Penrod, S. Improving the effectiveness of the Henderson instruction safeguard against unreliable eyewitness identification. *Psychol. Crime Law* **24**, 177–193 (2018).
58. Magnussen, S., Melinder, A., Stridbeck, U. & Raja, A. Q. Beliefs about factors affecting the reliability of eyewitness testimony: a comparison of judges, jurors and the general public. *Appl. Cogn. Psychol.* **24**, 122–133 (2010).
59. Martire, K. A. & Kemp, R. I. The impact of eyewitness expert evidence and judicial instruction on juror ability to evaluate eyewitness testimony. *Law Hum. Behav.* **33**, 225–236 (2009).
60. Safer, M. A. et al. Educating jurors about eyewitness testimony in criminal cases with circumstantial and forensic evidence. *Int. J. Law Psychiatry* **47**, 86–92 (2016).
61. Scurich, N. The differential effect of numeracy and anecdotes on the perceived fallibility of forensic science. *Psychiatry, Psychol. Law* **22**, 616–623 (2015).
62. Buhrmester, M., Kwang, T. & Gosling, S. D. Amazon's Mechanical Turk a new source of inexpensive, yet high-quality, data? *Perspect. Psychol. Sci.* **6**, 3–5 (2011).
63. Paolacci, G., Chandler, J. & Ipeirotis, P. Running experiments on Amazon Mechanical Turk. *Judgm. Decis. Mak.* **5**, 411–419 (2010).
64. Stewart, N. et al. The average laboratory samples a population of 7,300 Amazon Mechanical Turk workers. *Judgm. Decis. Mak.* **10**, 479–491 (2015).
65. Chandler, J., Mueller, P. & Paolacci, G. Nonnaïveté, among Amazon Mechanical Turk workers: consequences and solutions for behavioral researchers. *Behav. Res. Methods* **46**, 112–130 (2014).
66. Carpenter, B. et al. Stan: a probabilistic programming language. *J. Stat. Softw.* **76**, 1–32 (2017).
67. National Research Council *Strengthening Forensic Science in the United States: A Path Forward* (National Academies Press, 2009).
68. Laudan, L. & Allen, R. J. The devastating impact of prior crimes evidence and other myths of the criminal justice process. *J. Crim. Law Criminol.* **101**, 493–527 (2011).
69. Gross, S. R., O'Brien, B., Hu, C. & Kennedy, E. H. Rate of false conviction of criminal defendants who are sentenced to death. *Proc. Natl Acad. Sci. USA* **111**, 7230–7235 (2014).
70. Ngo, L. et al. Two distinct moral mechanisms for ascribing and denying intentionality. *Sci. Rep.* **5**, 17390 (2015).
71. Cole, S. A. Implementing counter-measures against confirmation bias in forensic science. *J. Appl. Res. Mem. Cogn.* **2**, 61–62 (2013).
72. Desmarais, S. L. & Read, J. D. After 30 years, what do we know about what jurors know? A meta-analytic review of lay knowledge regarding eyewitness factors. *Law Hum. Behav.* **35**, 200–210 (2011).
73. Ginther, M. R. et al. The language of mens rea. *Vanderbilt Law Rev.* **67**, 1327 (2014).
74. Koehler, J. J. & Meixner, J. B. in *The Psychology of Juries* (ed. Kovera, M. B.) 161–183 (American Psychological Association, Washington DC, 2017).
75. Scurich, N. What do experimental simulations tell us about the effect of neuro/genetic evidence on jurors? *J. Law. Biosci.* **5**, 204–207 (2018).
76. National Research Council *Identifying the Culprit: Assessing Eyewitness Identification* (National Academies Press, 2014); https://doi.org/10.17226/18891
77. Rouder, J. N. Optional stopping: No problem for Bayesians. *Psychon. Bull. Rev.* **21**, 301–308 (2014).
78. Pennington, N. & Hastie, R. Explanation-based decision making: effects of memory structure on judgment. *J. Exp. Psychol. Learn. Mem. Cogn.* **14**, 521–533 (1988).
79. Thompson, W., Black, J., Jain, A. & Kadane, J. *Forensic Science Assessments: A Quality and Gap Analysis—Latent Fingerprint Examination* (American Association for the Advancement of Science, 2017).
80. Gelman, A. et al. *Bayesian Data Analysis* (CRC Press, Boca Raton, 2013).

Acknowledgements

We thank D. Angel for valuable discussion throughout this project. A. Demasi, M. Rabil, M. Nerheim, P. Campos and A. Buras provided invaluable help recruiting subjects for these studies. We are especially grateful to the law students, prosecutors and other lawyers who volunteered their time as research subjects. This work was supported by an Incubator Award from the Duke Institute for Brain Sciences, NSF grant no. 1655445 and by a career development award from the NIH Big Data to Knowledge Program (grant no. K01-ES-025442 to J.M.P.). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

J.M.P. and J.H.P.S. conceived the project. All authors contributed to the task design. D.A.B., D.H.B. and J.A.G.S. wrote the crime scenarios and evidence modules. J.L., J.A.G.S. and J.M.P. designed and coded the online task, ran the mTurk experiments and processed the data. J.H.P.S. and D.H.B. recruited the legally trained subjects and administered the in-person sessions. J.M.P. wrote/designed the computational models. J.L., J.A.G.S. and J.M.P. analysed the data. J.H.P.S., J.M.P. and R.M.C. wrote the paper.

Competing Interests

D.A.B. and A.M. are partners in a trial consulting firm, whose practise includes civil litigation and criminal defence. A.M., J.H.P.S. and D.H.B. are attorneys, part of whose practises include criminal defence. J.H.P.S. works with both criminal defence attorneys and prosecutors. D.H.B. represents plaintiffs in civil cases and death row inmates in post-conviction criminal cases.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41562-018-0451-z.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to J.H.P.S.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

nature research

Corresponding author(s): J. H. Pate Skene

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
□	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	□	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
□	☒	A description of all covariates tested
□	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
□	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	□	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
□	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
□	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
□	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
□	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	We used custom R code in conjunction with the Stan Bayesian modeling language. All code used to run models, analyze, outputs, and produce figures is publicly available online at https://github.com/pearsonlab/legal.
Data analysis	We used custom R code in conjunction with the Stan Bayesian modeling language. All code used to run models, analyze, outputs, and produce figures is publicly available online at https://github.com/pearsonlab/legal.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Provide your data availability statement here.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

□ Life sciences ☒ Behavioural & social sciences □ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description

Study is a survey-based experiment with quantitative outcomes.

Research sample

Human participants were all healthy adults. Participants recruited online through Amazon Mechanical Turk indicated that they were at least 18 years old and registered with mTurk from United States addresses. Law students were second- and third-year students recruited from three U.S. law schools (Duke, Wake Forest, and University of Colorado-Boulder). Practicing prosecutors were members of the Lake County (Illinois) State's Attorney's office in and other practicing attorneys were recruited through email requests to participants in a continuing legal education conference sponsored by the Louisiana Bar Association.

Sampling strategy

Data were collected from a convenience sample of Amazon Mechanical Turk participants. Study postings included a brief description of the study and an amount paid to participants for successful completion. Law students were second- and third-year students recruited from three U.S. law schools (Duke, Wake Forest, and University of Colorado-Boulder). Practicing prosecutors were members of the Lake County (Illinois) State's Attorney's office in and other practicing attorneys were recruited through email requests to participants in a continuing legal education conference sponsored by the Louisiana Bar Association. For each of these legally-trained participant groups, completion of the study was purely opt-in.

Data collection

Data were collected via Qualtrics (for demographic data) and a custom web application designed by the research team. Participants viewed the study within a web browser. For Mechanical Turk participants, all data were collected remotely without the involvement of the research team. For law students, a member of the research team was present to introduce the study but not for data collection.

Timing

mTurk: Sept - Oct 2014, Sept 2014 - April 2015; Law students: March-April 2015; Lawyers May-June 2016; Prosecutors Oct 2016

Data exclusions

We excluded participants whose patterns of response either indicated a consistent failure to move the rating slider or unusually low time to complete the survey. Specifically, all responses from a subject were excluded if the data set met one of the following criteria:
1. Subject did not complete the entire task
2. Subject did not complete all 33 questions in test portion
3. Subject completed the test portion but not the demographics portion
4. Data suggested that subject did not participate in the task, or stopped participating in the task, due to repeated lack of change to default response values)
5. Unidentified and un-remediated coding errors that resulted in some subjects who were presented with the same question more than four times
6. Subject had participated in a previous version of the experiment (verified via Mechanical Turk user name).

Partial (repeat) data for an individual subject were excluded if the data set met the following criteria: Unidentified and un-remediated coding error that resulted in some subjects being presented with same question two, three, or four times. In those cases, responses after the first response were redacted.

Exclusion criteria were not established in the initial study design. We established the exclusion criteria after reviewing responses to the first round of data collection, each group comprising 50-100 subjects (Suppl. Table 2). Those criteria were then applied to all subsequent rounds of data collection.

Non-participation

This study was offered to all eligible Mechanical Turk workers. As a result, it is impossible to say how many declined. A small percentage of workers accepted the task but did not finish; data from these participants were excluded.

Randomization

Our studies did not use experimental and control groups. Each participant viewed each of 33 case "scenarios," each of which was randomly paired with one of 18 combinations of evidence. This pairing was done by sampling from the evidence combinations uniformly with replacement for each scenario.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	□ Unique biological materials
☒	□ Antibodies
☒	□ Eukaryotic cell lines
☒	□ Palaeontology
☒	□ Animals and other organisms
□	☒ Human research participants

Methods

n/a	Involved in the study
☒	□ ChIP-seq
☒	□ Flow cytometry
☒	□ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	See above.
Recruitment	See above.