Perceiving the arrow of time in autoregressive motion – Draft

Kristof Meding
University of Tübingen
Neural Information Processing Group
Tübingen, Germany
kristof.meding@uni-tuebingen.de

Dominik Janzing
Amazon Development Center
Amazon
Tübingen, Germany
janzind@amazon.com

Bernhard Schölkopf
Max-Planck-Institute for Intelligent Systems
Empirical Inference Department
Tübingen, Germany
bs@tuebingen.mpg.de

Felix A. Wichmann
University of Tübingen
Neural Information Processing Group
Tübingen, Germany
felix.wichmann@uni-tuebingen.de

Abstract

Understanding the principles of causal inference in the visual system has a long history at least since the seminal studies by Albert Michotte. Many cognitive and machine learning scientists believe that intelligent behavior requires agents to possess causal models of the world. Recent ML algorithms exploit the dependence structure of additive noise terms for inferring causal structures from observational data, e.g. to detect the direction of time series; the arrow of time. This raises the question whether the subtle asymmetries between the time directions can also be perceived by humans. Here we show that human observers can indeed discriminate forward and backward autoregressive motion with non-Gaussian additive independent noise, i.e. they appear sensitive to subtle asymmetries between the time directions. We employ a so-called frozen noise paradigm enabling us to compare human performance with three different algorithms on a trial-by-trial basis: A causal inference algorithm exploiting the dependence structure of additive noise terms, a neurally inspired network, and a Bayesian ideal observer model. Our results suggest that all human observers use similar cues or strategies to solve the arrow of time motion discrimination task, but the human algorithm is unique and significantly different from the three machine algorithms we compared it to. Additionally, our powerful frozen noise approach shows that although neural networks and ideal observer have remarkably similar performance they achieve this performance using different strategies.

1 Introduction

Discriminative convolutional neural networks (CNNs) have produced impressive results in machine learning, but certain striking failures of generalisation have been pointed out as well in terms of adversarial examples [1–3] or the recent findings of Geirhos and colleagues that CNNs show surprisingly large generalisation errors under image degradations [4, 5]. Many cognitive and machine learning scientists maintain that flexible and robust intelligent behaviour *in the real world* requires agents to possess generative or causal models of the world [6]. The importance of causality for cognitive science and psychology has long been recognized [7–16]. In visual perception, for example, it is fundamental to identify the causal structure in a visual scene: are objects moving or standing still, are some objects causing the movement of other objects [17–19], are the movements caused by

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

(intentional) actors or rather by forces of nature [20, 21]? On a cognitive rather than perceptual level progress has been made to understand how we intuitively understand physics [22], how humans learn causal structures from data [10, 7, 8, 13] and on human causal inference via counterfactual reasoning [16, 23].

Much less research has explored whether the earlier, perceptual and unconscious—cognitively impenetrable [24]—processing stages in humans possess already causal inference algorithms, see Danks [25] for a recent overview on the relationship between causal perception, causal inference and causal reasoning. Rolfs et al. [26] found evidence for perceptual adaptation to causality, thus arguing that the perceptual system already possesses mechanisms tuned to "causal features" in the visual input (but c.f. for a critique of the paper on methodological grounds [27]). More recently it was shown, using the continuous flash suppression paradigm, that simple Michotte style launching-events enter awareness faster when they are perceived as causal events, again suggesting that rather early, perceptual and pre-conscious processes may already be tuned to "causal features" [28].

Recently there has been considerable progress in understanding causal inference by approaching it as a machine learning problem [29–32]. In the last two decades algorithms for causal inference with different approaches have been suggested. Based on the language of graphical models and structural equation models, the "classical approaches" infer the directed acyclic graph (DAG) formalizing the causal relations from the observed conditional statistical (in)dependences subject to causal Markov condition and causal faithfulness [29, 30]. After about 2004, several other approaches were suggested that infer causal DAGs using properties of distributions other than conditional independence. These approaches also consider DAGs that consist of two variables only (in which case conditional independence testing is futile), i.e., to decide what is cause and what is effect, see chapter 4 of [32] for an overview. It was shown that one can still infer the structure if one is willing to place restrictions on the action of the noise disturbances, specifically, that it is additive and independent, and that either the noise is non-Gaussian or the functions are nonlinear [33–36]. These methods have also been applied to determine the causal direction of time series by fitting autoregressive models, i.e., by predicting future from past, and examining the noise terms [37–39].

Investigation of the arrow of time in causal learning was motivated by its role in physics [40, 41, 37], since it can be shown that the time asymmetry based on the independence of noise can be explained by the usual thermodynamic arrow of time [42] and that recent approaches to causal inference are thus linked to statistical physics [43]. Pickup et al. showed that the independence of noise can be employed to detect the arrow of time in real world YouTube videos, without semantic or cognitive knowledge about the visual world [39]. Recently it was shown that also neural networks can infer the arrow of time from movies alone [44], suggesting that even low-level motion information in the video contains information about the arrow of time.

Clearly, humans can perceive the arrow of time in settings where semantic information or world knowledge is available. In a famous movie by the Lumière brothers, a wall falls over, subsequently shown backwards to illustrate the perceptual contrast.[1] Similarly, humans can perceive the arrow of time if there is a clear non-stationarity in the data, or a directionality due to a perceivable increase in entropy, e.g. if we observe an explosion. However, ML causality methods can also infer the arrow of time in cases that at first sight appear hard, i.e. where the marginals are the same in both motion directions and the setting is stationary. For humans, in contrast, the perception of the arrow of time in such settings is unclear. Although it is well established that humans are sensitive to higher-order regularities in the *spatial* statistics of static natural images [45], for motion sequences or motion discrimination analogous results have not yet been established. It was even recently shown—at least when assessing the motion direction of random dot kinematograms (RDKs)—that humans appear only sensitive to the mean and variance of the displacement angles but were insensitive to skewness and kurtosis [46]. Thus, for RDKs, and unlike in the case of static spatial structure, the human visual system appears insensitive to higher-order statistics. Causal dependency algorithms, however, in the linear case crucially rely on non-Gaussianity of additive noise, for which kurtosis and additional non-zero higher-order moments are a measure.

Thus we investigated whether the human visual system is sensitive to dependencies in the motion of a single disk. Furthermore, we investigate in depth the relationship between the abilities of different machine learning algorithms: a Residual Dependence based algorithm, a Neural Network and a Bayesian Ideal Observer. We show, first, that human observers can indeed discriminate the arrow of

[1] `https://www.youtube.com/watch?v=W_bB0TVTwg8`

time in autoregressive (AR) motion with non-Gaussian additive independent noise, i.e. they appear sensitive to subtle time reflection asymmetries. Second, we show that humans are remarkably unique and efficient in this task, requiring only a short motion sequence to identify the direction of the time series. Third, we show that the ideal observer algorithm and the neural network both achieve "super-human"—and quantitatively very similar—performance, but the frozen noise paradigm we employed shows that both algorithms use different cues or strategies.

2 Methods

Here we provide the minimum information necessary to understand our experiments and results. We refer to the supplementary material for detailed explanations and all information needed to allow all experiments to be reproduced.

2.1 The arrow of time: Causal and anti-causal time series

We constructed time series from a generative additive noise model:

$$x_t = 0.05 \cdot x_{t-4} + 0.1 \cdot x_{t-3} + 0.2 \cdot x_{t-2} + 0.4 \cdot x_{t-1} + \epsilon_t$$

The noise ϵ_t is independent from all previous states $x_{t-1}, x_{t-2}, ...$ Clearly, future states $x_t, x_{t+1}, x_{t+2}, ...$ are dependent on ϵ_t since ϵ_t influences them (the arrow of time in this setting). This is true for all types of noise distributions for ϵ_t, however, the direction is not detectable for Gaussian noise in a linear time series since a linear Gaussian time series can be modeled in the forward and backward direction with independent noise terms. For non-Gaussian noise, however, this is not true: it is not possible to fit a time series in the backward direction with independent noise terms [37].

Multiple algorithmic ways exist to detect the direction of such a time series based on this dependence structure. We describe them in section 2.3. Note that we can use the case with the Gaussian distribution for ϵ_t as "sanity check" to test our psychophysical experiment as well as our algorithms: neither humans nor algorithms should be able to identify the direction with Gaussian noise.

Throughout we use time series for which the additive noise component ϵ_t is distributed according to $\epsilon_t \sim sgn(Y) \cdot |Y|^r$, with Y Gaussian distributed. We choose the exponent r in the range of $0.1 - 6$. This yields noise which is either Bimodal ($r < 1$) or peaked Super-Gaussian ($r > 1$). The closer the value of r is to 1, the more Gaussian ϵ_t becomes. An Exponent $r = 1$ yields Gaussian distributed noise. The noise parameterization with exponent r has the advantage that the non-Gaussianity of the time series can be precisely controlled with one parameter. Additionally, we choose a single smoothed Uniform distribution with tails extending to $\pm\infty$. In total 16 time series were used in our experiment, seven with Super-Gaussian additive noise, seven with Bimodal additive noise, one with smoothed Uniform and one with Gaussian additive noise. All noise distributions had mean 0 and standard deviation of 44.72 pixels on screen (1,13 cm). These values ensure, in practice, that the time series is bounded to the range of possible coordinates of the monitor used in our experiment. Time series in the true time direction are in the following denoted as causal time series, and time series which are flipped along the temporal axis are denoted as anti-causal. Movies of the stimuli are presented in the supplementary material.

2.2 Psychophysical Experiment

We tested if humans have the ability to discriminate causal from anti-causal time series in a psychophysical experiment. Observers saw a white random dot moving across the horizontal axis on a computer screen. The dot position followed a linear non-Gaussian time series with additive noise described as above. Observers had to press a button whether they saw the moving dot belonging to the green (causal) or to red (anti-causal) category—observers were unaware that the difference between the categories was a time-reversal; they were given a cover story to identify harmless from dangerous bacteria based on their motion. We hypothesized that humans are better in classifying very strong non-Gaussian time series as algorithms do [37]. Thus we began by training subjects with easily classifiable noise and made the time series progressively more difficult (making r approach 1.0). Human observers should be able to use the same cue for different intensities of the Bimodal or Super-Gaussian noise. The discrimination task is rather difficult and we screened participants based on their performance in what we considered "easy conditions" with $r = 6, 4, 2$ (Super-Gaussian) and

$r = 0.1, 0.3, 0.5$ (Bimodal). Participants had to achieve at least 67.5% in these blocks (40 trials) to be significantly different from chance level and to participate further in our experiment. Seven of our 17 naive observers failed to reach the criterion. We provide detailed information in the supplementary material A.2.2 why we think this does not influence our overall results about human performance

Ten naive observers participated successfully in the first experiment (6 female, 4 male mean, mean age = 24 yrs, std = 2.5 yrs). All subjects received monetary compensation. The observers were tested on time series with all 16 noise distributions. For Bimodal and Super-Gaussian noise observers progressed from easy to difficult noise. Each observer classified every of the 16 noise distributions 40 times, 640 trials in total per observer and it took each observer four hours to complete the first experiment.

The first experiment assessed how well observers were able to discriminate forward and backward AR motion sequences as a function of the degree of non-Gaussianity of the additive noise, i.e. to see the arrow of time. Our second experiment aimed to investigate both human and algorithmic strategies for the detection of the arrow of time. In this experiment the noise was randomly samples for all subjects. To this end all subjects were tested on exactly the same time series: the so-called frozen noise paradigm often successfully employed in auditory psychophysics [47, 48] This experimental technique allows to examine inter-subject or subject-algorithmic correlation and consistency. In the second experiment we only used a single noise, Bimodal noise with exponent $r = 0.5$. The length of the motion sequence—and thus the viewing time—was reduced progressively from the initial 100 time-points to finally only 2 time points (100, 50, 25, 20, 16, 12, 8, 4, 2). Participants classified 40 trials for each sequence length yielding in total 360 trials per observer. Similar to experiment one the task got more difficult as the experiment progressed. Four of the best observers in the previous experiment participated in this experiment (2 male, 2 female, age =22.5 yrs, std = 2.3 yrs). The experiment lasted 1.5 hours per observer.

2.3 Algorithms for causal inference

One central aim of ours is to compare the abilities of humans and algorithms to detect the arrow of time. We compared the performance of our human observers to three different algorithms: First, an algorithm which directly exploits the residual dependence structure (ResDep). Second, a neurally inspired network and, third, a Bayesian ideal observer algorithm.

The ResDep algorithm proposed by Jonas Peters et al. [37] uses directly the residual dependence structure of ϵ_t to the value x_{t-1}. The algorithm fits an autoregressive model to the time series and a series flipped along the time dimension. Subsequently an independence test is performed between fitted residuals and data points. The direction is decided using the Hilbert-Schmidt Independence Criterion test. The true time direction maximizes the independence-score between residuals and data points.

The second algorithm was a (simple) neurally inspired network [49, 50]. The network consisted of one convolution layer, followed by a batch normalization layer, a ReLU-layer and a fully connected layer (see A.2.4 in the appendix for further details). For each noise distribution the network was trained with 30000 time series. We used the Adam optimizer with an initial learning rate of 0.01. The network was trained for a maximum of 30 epochs. Both the ResDep algorithm and the neural network has full temporal memory since we input the full time series at the first step.

While the neural network gets as input the full time series and thus has perfect temporal memory, we can contrast this algorithm with one based on Bayes statistics. In the vision literature this is often done in an ideal observer framework [51]. An ideal observer analysis is a statistical framework which provides the upper limit of performance *given a set of constraints* since the ideal observer has perfect knowledge about the underlying model and its constraints.

We calculated the probability of the direction d given the data $X = (x_1, x_2, ..., x_N)$ using Bayes rule:

$$p(d|X) = \frac{p(X|d) \cdot p(d)}{p(X)} = \frac{\Pi_{t=1}^{N} p(x_t|x_{t-1}, x_{t-2}, ..., x_1, d) \cdot p(d)}{p(X)}.$$

If we consider only stationary and stable time series of order 4—as in our experiments—the terms in the numerator become $p(x_t|x_{t-1}, x_{t-2}, x_{t-3}, x_{t-4})$ for the forward time series. This term corresponds exactly to the chosen noise distribution. We compare this expression in the forward and backward direction and choose the direction for which the corresponding probability is larger. This method is very similar to calculating the Bayes Factor. See section A.2.4 in the appendix for a detailed explanation.

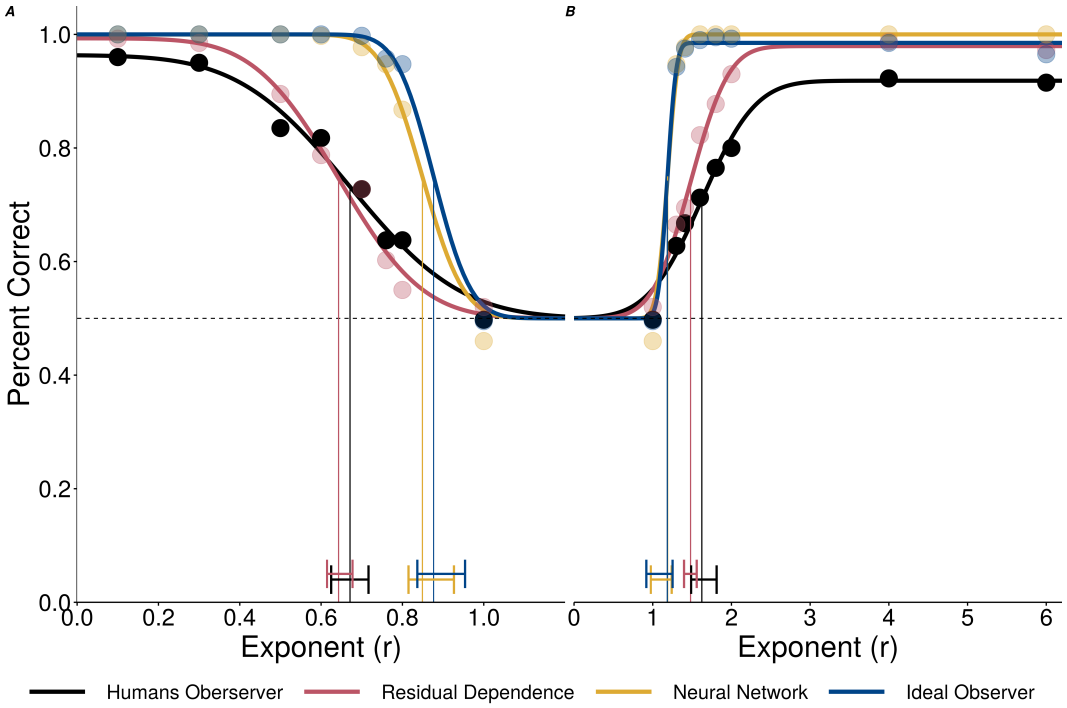

Figure 1: Psychometric Function for Bimodal noise (A) and Super-Gaussian noise (B). The black dots represent the human accuracy for different exponents, pooled over all 10 subjects. The psychometric functions are fitted with cumulative Gaussian distributions. Performance gets worse towards an exponent of 1 which corresponds to the non-identifiable Gaussian noise case. The horizontal line marks the width, the scaled 75% threshold for the different fits. Whiskers show 95% Credible Intervals (CI) for the threshold.

3 Results

The following psychometric functions and Bayesian Credible Intervals (CI) were calculated with the Beta Binomial Model in Psignifit 4 [52]. Figure 1 shows the main result of Experiment one. The black psychometric functions show the human data (pooled across all ten observers) and the coloured curves results for the algorithms on the same time series seen by our human observers: ResDep in red, the neural network in yellow, the ideal observer in blue. Data for single observers are shown in Figure A.4. All individual psychometric functions can be found in figure A.5 and A.6, and the thresholds with credible intervals in table 4. On the one hand, humans can indeed detect the direction of a time series for Super-Gaussian and Bimodal noise with thresholds $\mathrm{r} = 0.67, 95\%$ CI $[0.62, 0.72]$ (Bimodal) and $\mathrm{r} = 1.62, 95\%$ CI $[1.45, 1.81]$ (Super-Gaussian). The ResDep algorithm, on the other hand, performs similar to humans with Bimodal noise (threshold $\mathrm{r} = 0.64, 95\%$ CI $[0.61, 0.68]$) and, perhaps, marginally better with Super-Gaussian noise (threshold $\mathrm{r} = 1.48, 95\%$ CI $[1.4, 1.56]$)[2]. Algorithmic performance of the neural network and the ideal observer is superior to human and ResDep performance and both algorithms show remarkably similar results. A detailed analysis of the neural network can be found in A.4. Thresholds for the exponents of the Neural Network are $\mathrm{r} = 0.85, 95\%$ CI $[0.82, 0.93]$ and $\mathrm{r} = 1.19, 95\%$ CI $[0.98, 1.24]$ and for the ideal observer $\mathrm{r} = 0.88, 95\%$ CI $[0.83, 0.95]$ and $\mathrm{r} = 1.18, 95\%$ CI $[0.92, 1.25]$.
The parameterization with exponent r is somewhat arbitrary and we tested in figure A.10 other distant scales (Kullback-Leibler divergence, Jensen-Shannon divergence, Kolmogorov-Smirnov distance and normalized exponents).

The result for the smoothed Uniform noise were much more diverse—remember that there is only a single smoothed Uniform distribution with zero mean and the same variance as all other noise distributions we used: The average human accuracy was 50% (chance performance), for ResDep

[2]The best three human observers for Super-Gaussian noise had thresholds of $r = 1.32, 1.36, 1.38$—at least as sensitive as ResDep.

70%, for the neural network 96% and for the ideal observer 97%; we discuss these results in the next section.
From Figures 1 and the block by block comparison in A.11 it appears as if human observers may use an internal algorithm similar to ResDep (top left panel in A.11), and the neural network may have learned a strategy mimicking that of the ideal observer (bottom right panel in A.11).

The frozen noise paradigm described above in section 2.2 and used in our experiment 2 allows us to investigate this question in a much more stringent way: All human observers and the algorithms classified exactly the same time series—they were not only drawn from the same distribution but the very same time series. In addition, in experiment 2 we explored how human observers and algorithms cope with shorter time series (Bimodal noise, $r = 0.5$ fixed throughout the experiment). This, too, may offer a way to distinguish human observers and algorithms from each other.

Figure 2 shows the results for experiment 2. Plotting conventions as in Figure 1: The black psychometric function shows human data (pooled across all four observers) and the coloured curves results for the algorithms on exactly the same time series seen by our human observers: ResDep in red, the neural network in yellow, the ideal observer in blue. Individual psychometric functions of the four human observers are shown in Figure A.12. The neural network was exactly trained as in experiment one with the exception that we shrink the size of the convolutional layer for very short time series, see A.2.4 for details. Human observers are able to detect the direction of time series even for rather short time series, with a threshold of about 17.76, 95% CI [14.40,22.44] time points. The results are even more impressive if we exclude subject BW—who told us after the experiment that he had been not fully attentive during the experiment: the threshold drops to 15.17 time points, 95% CI [11.51,19.18], see figure A.13 and figure A.14 in the appendix. In this respect, humans clearly outperform the ResDep algorithm which requires 42.67, 95% CI [28.88,58.86] time points for 75% correct discrimination. The neural network with a threshold of 8.13, 95% CI [-1.85,12.52] time points and the ideal observer algorithms with a threshold of 7.73, 95% CI [-0.71,11.16] again show similar performance and are again superior to that of human observers and ResDep. Please note, however, that the somewhat poor performance of ResDep may not (only) reflect its intrinsic inferiority but may in part be due to the difficulty of fitting short time series. ResDep relies on the ARMA method in MATLAB; ResDep is effectively guessing for time series shorter or equal than 8 time points. Also, the ideal observer has intrinsic problems with short time series since our underlying assumptions for the approximation does not hold anymore, see A.2.4 in the supplementary material for further details. The frozen noise method allows us to compare observer consistency within observers and consistency between humans and algorithms. If subject 1 has for a given block an accuracy p_1 and subject 2 has for the same block an accuracy p_2, then we would expect for independent binomial observers a fraction of $p_1 \cdot p_2 + (1-p_1) \cdot (1-p_2)$ equally answered ("consistent") trials. This fraction of expected consistency is compared to the number of actually equally answered trials per block. If the observed proportion is significantly higher than expected, this provides evidence that subjects 1 and 2—be them two humans, two algorithms or a human and an algorithm—are not independent, which in turns indicates that they rely on similar processing strategies or at least use similar stimulus information.
Figure 3 shows this comparison for humans and algorithms, with the expected consistency shown on the x-axis, plotted on the observed consistency on the y-axis. Comparing human observers to each other (top left panel in Figure 3, we see that humans tend to have more similarities than expected from independent observers. (The shaded ellipsoidal regions indicate the confidence regions around the null hypothesis that they are independent given the amount of data.) The first column in Figure 3) strongly suggests that humans observers use a strategy or internal algorithm independent from all three of our ML algorithms. Furthermore, the graph shows that all algorithms show only a consistency consistent with them being independent. (Because we can generate more data for the algorithmic comparisons, we confirmed this using many more trials, reaching the same conclusion, see Figure A.15.)

4 Conclusion

Our frozen noise paradigm shows that ideal observer and the neural network have unique strategies. Even if we use more data point in figure A.15 we see only a small effect of similarity. One could argue that we do not find an agreement of the ideal observer and neural network due to the intrinsic problem of the ideal observer algorithm for short time series. But even if we redo the frozen noise paradigm using long sequences but varying the exponent—thus rendering the sequences difficult not by shortening them but by making the noise more Gaussian—we again see only a minor effect, see figure A.16. The ideal observer and the neural network use different, albeit equally successful,

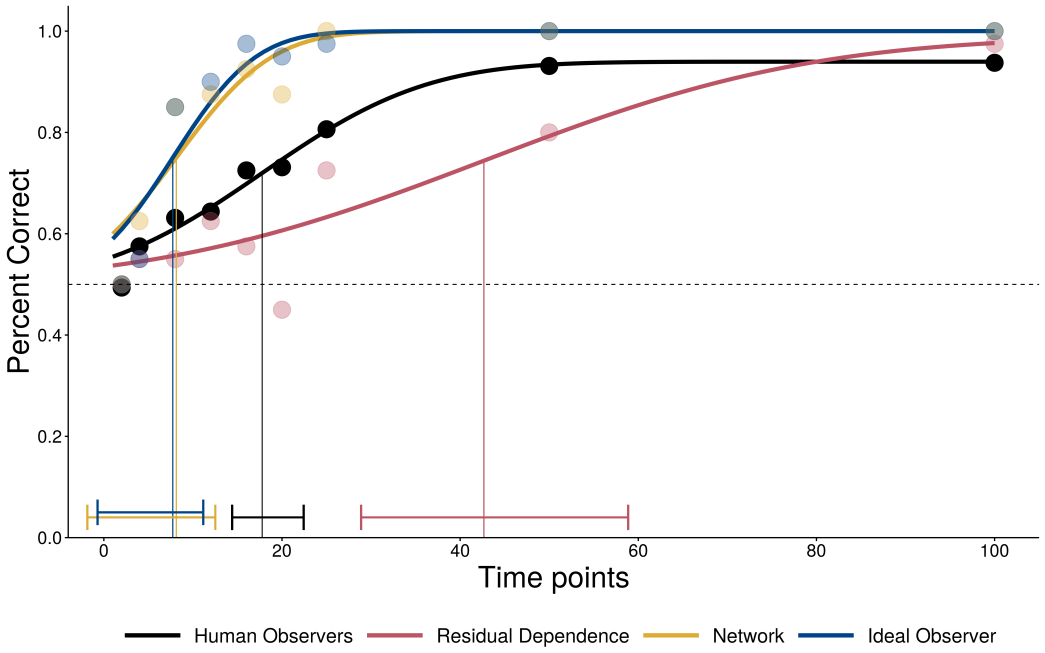

Figure 2: Performance of humans and algorithms for time series of different lengths, plot conventions are as in figure 1

.

strategies.

Despite the fact that, on the one hand human observers and ResDep and on the other hand the neural network and the ideal observer, show roughly the same performance in experiment 1, the frozen noise paradigm in experiment 2 allows us to conclude that they actually all use independent strategies. In particular, human observers do not use a ResDep dependency algorithm, and neither do they use an ideal Bayesian probability calculation—especially the latter is a popular notion in visual perception and the cognitive sciences.

Another main outcome of our study is how remarkably efficient the unique strategy of the visual system is: Our observers only needed 17.76 95% CI [14.40,22.44] time points (15.17, 95% CI [11.51,19.18] if we exclude one somewhat poorer performing observer) for 75% correct discrimination of the forward or backward played AR motion sequences (the arrow of time). They require fewer data samples than a successful ML algorithm for causal inference (ResDep; with the caveat regarding implementation mentioned above. A different implementation of the ResDep ideas may perform better). Performance approached that of the ideal observer that knows the underlying statistics perfectly, i.e., the order of the AR process, the AR coefficients, the variance and exponent of the noise of the time series. We deem it unlikely that the human observers could extract these parameters from visual input alone, let alone for the very short sequences.

When performing demanding psychophysical experiments with human observers there is always the question of learning—are we really reporting and interpreting stationary performance? In causal inference in cognitive science structure learning from data is an important topic (e.g. dynamical causal learning [10]). However, in our experiments observers were able to do the motion discrimination after a few training trials. More importantly, the accuracy in the first and second half of every block was very similar. Average performance in experiment 1 pooled for all subjects and across all noise distributions was 82% for trials 1-20 and 80% for trials 20-40; 64%/63% in experiment 2. This strongly suggests that our data are not contaminated by learning effects during our experiments.

One puzzle we are unable to resolve is why our human observers typically failed to reach above chance performance with the smoothed Uniform distribution: performance for smoothed Uniform was at 50% across all observers. From a psychophysical point of view the smoothed Uniform condition was more difficult by experimental design: Observers could not start with easy smoothed Uniform noise since there was no free parameter. On the other hand, Bimodal and smoothed Uniform

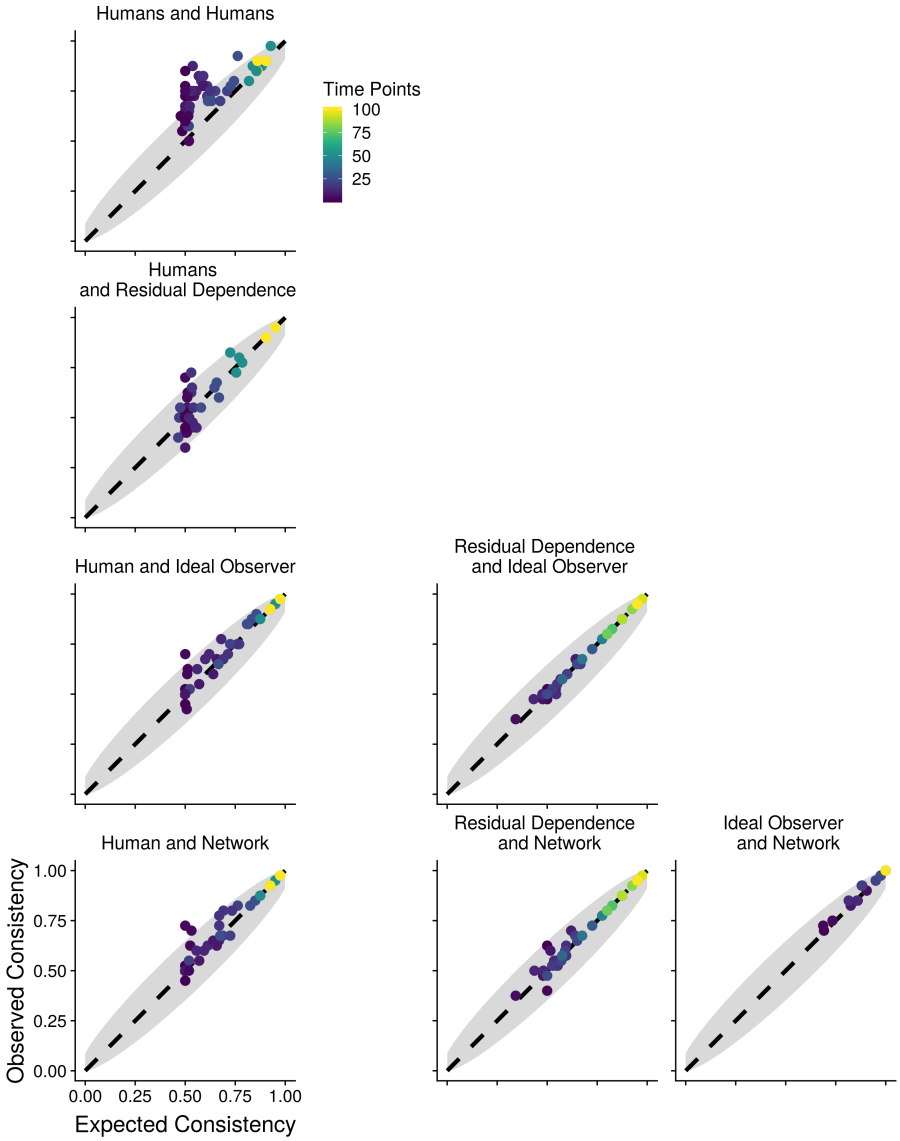

Figure 3: Human observer consistency and observer-algorithmic consistency for the frozen noise paradigm. The x-axis shows the expected proportion of equally answered trials under the assumption of independent observers or algorithms. The y-axis shows the actual observed number of equally answered trials in the experiment. The shaded area shows a 95% confidence interval calculated based on the Wilson score interval [53]. Colour codes the number of time points. We used in the algorithm-algorithm comparison not only time series with lengths from the experiment but also a finer grid: 10-30 time points with spacing 1 and 35-100 with spacing 5.

distributions have a similar dependence structure, see Figure A.1. We expected that at least those observers that were already trained on Bimodal noise should be able to detect the direction of the smoothed Uniform time series—however, that was not the case. Only observer LL achieved an accuracy above 65%. (The JS-Divergence of the smoothed Uniform distribution corresponds to a Bimodal exponent of 0.73. As we can see from Figure 1 we expect around 65% performance, in line with LL's performance.) The surprising difficulty of the smoothed Uniform distribution should help constrain which strategy or algorithm was used by our human observers during our experiments. Recent advances in causal inference have been strongly driven by human intuition about how the shape of joint distribution indicates causal directions [54, 33]. This line of argument, together with our experimental results, suggests that many of the human abilities regarding the recognition of causal and time asymmetries are not known yet.

Ever since Albert Michotte performed his studies there is the question whether causal inference may under certain circumstances already be a perceptual rather than a cognitive ability [26, 28]. In our experiment observers were able to discriminate very subtle temporal asymmetries, similar to the remarkable sensitivity to higher-order spatial dependencies in patches of natural images [45]. To us this hints at an early, perceptual locus in our experiments.

References

[1] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv*, 1312.6199v4:1–10, 2014.

[2] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016.

[3] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, pages 506–519. ACM, 2017.

[4] Robert Geirhos, Carlos R Medina Temme, Jonas Rauber, Heiko H Schütt, Matthias Bethge, and Felix A Wichmann. Generalisation in humans and deep neural networks. *Advances in Neural Information Processing Systems*, 31:7549–7561, 2018.

[5] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *International Conference on Learning Representations (ICLR)*, 2019.

[6] Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40:e253, 2017.

[7] Patricia W Cheng. From covariation to causation: a causal power theory. *Psychological review*, 104(2): 367–405, 1997.

[8] Joshua B Tenenbaum and Thomas L Griffiths. Structure learning in human causal induction. In *Advances in neural information processing systems*, pages 59–65, 2001.

[9] Tamar Kushnir, Alison Gopnik, Laura Schulz, and David Danks. Inferring hidden causes. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 25, 2003.

[10] David Danks, Thomas L Griffiths, and Joshua B Tenenbaum. Dynamical causal learning. In *Advances in neural information processing systems*, pages 83–90, 2003.

[11] Mark Steyvers, Joshua B Tenenbaum, Eric-Jan Wagenmakers, and Ben Blum. Inferring causal networks from observations and interventions. *Cognitive science*, 27(3):453–489, 2003.

[12] Alison Gopnik, Clark Glymour, David M Sobel, Laura E Schulz, Tamar Kushnir, and David Danks. A theory of causal learning in children: causal maps and bayes nets. *Psychological review*, 111(1):3–32, 2004.

[13] Thomas L Griffiths and Joshua B Tenenbaum. Structure and strength in causal induction. *Cognitive psychology*, 51(4):334–384, 2005.

[14] Alison Gopnik and Laura Schulz. *Causal learning: Psychology, philosophy, and computation*. Oxford University Press, 2007.

[15] Noah D Goodman, Tomer D Ullman, and Joshua B Tenenbaum. Learning a theory of causality. *Psychological review*, 118(1):110, 2011.

[16] David A Lagnado, Tobias Gerstenberg, and Ro'i Zultan. Causal responsibility and counterfactuals. *Cognitive science*, 37(6):1036–1073, 2013.

[17] Albert Michotte. *The perception of causality*. Oxford, England: Basic Books, 1963.

[18] Alan M Leslie and Stephanie Keeble. Do six-month-old infants perceive causality? *Cognition*, 25(3): 265–288, 1987.

[19] Lance J Rips. Causation from perception. *Perspectives on Psychological Science*, 6(1):77–97, 2011.

[20] Fritz Heider and Marianne Simmel. An experimental study of apparent behavior. *The American journal of psychology*, 57(2):243–259, 1944.

[21] Brian J Scholl and Tao Gao. Perceiving animacy and intentionality: Visual processing or higher-level judgment. *Social perception: Detection and interpretation of animacy, agency, and intention*, pages 197—-230, 2013.

[22] James R Kubricht, Keith J Holyoak, and Hongjing Lu. Intuitive physics: Current research and controversies. *Trends in cognitive sciences*, 21(10):749–759, 2017.

[23] T. Gerstenberg, N. D. Goodman, D. A. Lagnado, and J. B. Tenenbaum. How, whether, why: Causal judgments as counterfactual contrasts. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, Jennings Matlock, T., C. D., and P. P. Maglio, editors, *Proceedings of the 37th Annual Conference of the Cognitive Science Society*, pages 782–787, Austin, TX, 2015. Cognitive Science Society.

[24] Jerry A Fodor. *The Modularity of Mind*. MIT Press, 1983.

[25] David Danks. The psychology of causal perception and reasoning. In H. Beebee, C. Hitchcock, and P. Menzies, editors, *Oxford handbook of causation*, pages 447–470. Oxford University Press, 2010.

[26] Martin Rolfs, Michael Dambacher, and Patrick Cavanagh. Visual adaptation of the perception of causality. *Current Biology*, 23(3):250–254, 2013.

[27] Regan M. Gallagher and Derek H. Arnold. Comparing the aftereffects of motion and causality across visual co-ordinates. *bioRxiv*, 2018.

[28] Pieter Moors, Johan Wagemans, and Lee de-Wit. Causal events enter awareness faster than non-causal events. *PeerJ*, 5:e2932, 2017.

[29] Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2000.

[30] Judea Pearl. *Causality: models, reasoning, and inference*. Cambridge University Press, 2000.

[31] Daniel Malinsky and David Danks. Causal discovery algorithms: A practical guide. *Philosophy Compass*, 13(1):e12470, 2018.

[32] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. MIT press, 2017.

[33] Patrik O Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in neural information processing systems*, pages 689–696, 2009.

[34] Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, and Antti Kerminen. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(Oct):2003–2030, 2006.

[35] Kun Zhang and Aapo Hyvärinen. On the identifiability of the post-nonlinear causal model. In *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*, pages 647–655. AUAI Press, 2009.

[36] Jonas Peters, Joris M Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *The Journal of Machine Learning Research*, 15(Jun):2009–2053, 2014.

[37] Jonas Peters, Dominik Janzing, Arthur Gretton, and Bernhard Schölkopf. Detecting the direction of causal time series. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 801–808, 2009.

[38] Stefan Bauer, Bernhard Schölkopf, and Jonas Peters. The arrow of time in multivariate time series. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, pages 2043–2051, 2016.

[39] Lyndsey C Pickup, Zheng Pan, Donglai Wei, YiChang Shih, Changshui Zhang, Andrew Zisserman, Bernhard Schölkopf, and William T Freeman. Seeing the arrow of time. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2035–2042, 2014.

[40] Hans Reichenbach. *The direction of time*, volume 65. University of California Press, Berkeley, USA, 1956.

[41] Huw Price. *Time's arrow & Archimedes' point: new directions for the physics of time*. Oxford University Press, 1997.

[42] D. Janzing. On the entropy production of time series with unidirectional linearity. *Journ. Stat. Phys.*, 138: 767–779, 2010.

[43] D. Janzing, R. Chaves, and B. Schölkopf. Algorithmic independence of initial condition and dynamical law in thermodynamics and causal inference. *New Journal of Physics*, 18(093052):1–13, 2016.

[44] D. Wei, J. Lim, A. Zisserman, and W. T. Freeman. Learning and using the arrow of time. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[45] Holly E. Gerhard, Felix A. Wichmann, and Matthias Bethge. How sensitive is the human visual system to the local statistics of natural images? *PLOS Computational Biology*, 9(1):1–15, 01 2013.

[46] Michael L Waskom, Janeen Asfour, and Roozbeh Kiani. Perceptual insensitivity to higher-order statistical moments of coherent random dot motion. *Journal of vision*, 18(6):9–9, 2018.

[47] Newman Guttman and Bela Julesz. Lower limits of auditory periodicity analysis. *The Journal of the Acoustical Society of America*, 35(4):610–610, 1963.

[48] Trevor R Agus, Simon J Thorpe, and Daniel Pressnitzer. Rapid formation of robust auditory memories: insights from noise. *Neuron*, 66(4):610–618, 2010.

[49] Robbe LT Goris, Tom Putzeys, Johan Wagemans, and Felix A Wichmann. A neural population model for visual pattern detection. *Psychological review*, 120(3):472–496, 2013.

[50] Heiko H Schütt and Felix A Wichmann. An image-computable spatial vision model. *Journal of Vision*, 17 (12):12, 1–35, 2017.

[51] Wilson S Geisler. Ideal observer analysis. *The visual neurosciences*, 10(7):12–12, 2003.

[52] Heiko H. Schütt, Stefan Harmeling, Jakob H. Macke, and Felix A. Wichmann. Painfree and accurate bayesian estimation of psychometric functions for (potentially) overdispersed data. *Vision Research*, 122: 105 – 123, 2016.

[53] Edwin B Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927.

[54] X. Sun, D. Janzing, and B. Schölkopf. Causal inference by choosing graphs with most plausible Markov kernels. In *Proceedings of the 9th International Symposium on Artificial Intelligence and Mathematics*, pages 1–11, Fort Lauderdale, FL, 2006.

[55] David H Brainard. The psychophysics toolbox. *Spatial vision*, 10:433–436, 1997.

[56] M Kleiner, D Brainard, Denis Pelli, A Ingling, R Murray, and C Broussard. What's new in psychtoolbox-3. *Perception*, 36(14):1–16, 2007.

[57] H. Akaike. Information theory and an extension of the maximum likelihood principle. Second International Symposium on Information Theory, Tsahkadsor, Armenia (Petrov, B.N. and Csaki, F., eds), 267-281., 1973.

A Supplementary Material

Section A.1 and A.2 provide in depth explanation and all information necessary for reproducing our experiments. Section A.3 contains an analysis of the neural network. Section A.4-A.10 show additional figures.

A.1 Generation of time series

We constructed time series with different noise types. Table 1 lists all distributions used in constructing time series. Figure A.1 shows the distributions' densities and the dependence of residuals in forward and backward direction for a fourth-order time series. This independence is used by the algorithm proposed by Peters et al. [37].

The (forward) time series are constructed by the following rule

$$x_t = 0.05 \cdot x_{t-4} + 0.1 \cdot x_{t-3} + 0.2 \cdot x_{t-2} + 0.4 \cdot x_{t-1} + \epsilon_t. \tag{1}$$

The first four values of x_t were set to zero and the consecutive 400 time points are drooped in order to make time series stationary. The mean of all noise distributions was set to 0 and the standard deviation to 44.72 pixels on screen (1,13 cm). These values ensure that the time series is bounded to the range of possible coordinates of the monitor used in our experiment. Backward time series were constructed in the way that we first constructed new forward time series and then flip series along the time axis

In the Super-Gaussian and Bimodal case the noise was created according to the following rule:

$$\epsilon_t = sgn(Y) \cdot |Y|^r, \tag{2}$$

while $Y \sim \mathcal{N}(\mu, \sigma^2)$.We calculate, with the change of variable technique, the density of the noise as

$$p(e_t) = \frac{1}{r} \cdot |\epsilon_t|^{\frac{1}{r}-1} \frac{1}{\sqrt{2\pi}\sigma} \cdot e^{-\frac{\epsilon_t^{\frac{2}{r}}}{2\sigma^2}}. \tag{3}$$

The variance is calculated as

$$Var(\epsilon_t) = \frac{2^r}{\sqrt{\pi}} \cdot \sigma^{2r} * \Gamma(r + \frac{1}{2}). \tag{4}$$

This variance of the noise depends only on σ and the exponent r. While fixing the exponent, the free parameters σ can be used to set the variance of the noise distribution to 2000.

Table 1: Noise distributions used for constructing time series and the corresponding exponents.

Distribution	Parameter (r)
Super-Gaussian	6
Super-Gaussian	4
Super-Gaussian	2
Super-Gaussian	1.8
Super-Gaussian	1.6
Super-Gaussian	1.41
Super-Gaussian	1.3
Bimodal	0.1
Bimodal	0.3
Bimodal	0.5
Bimodal	0.6
Bimodal	0.7
Bimodal	0.76
Bimodal	0.8
Gaussian	1
smoothed Uniform	-

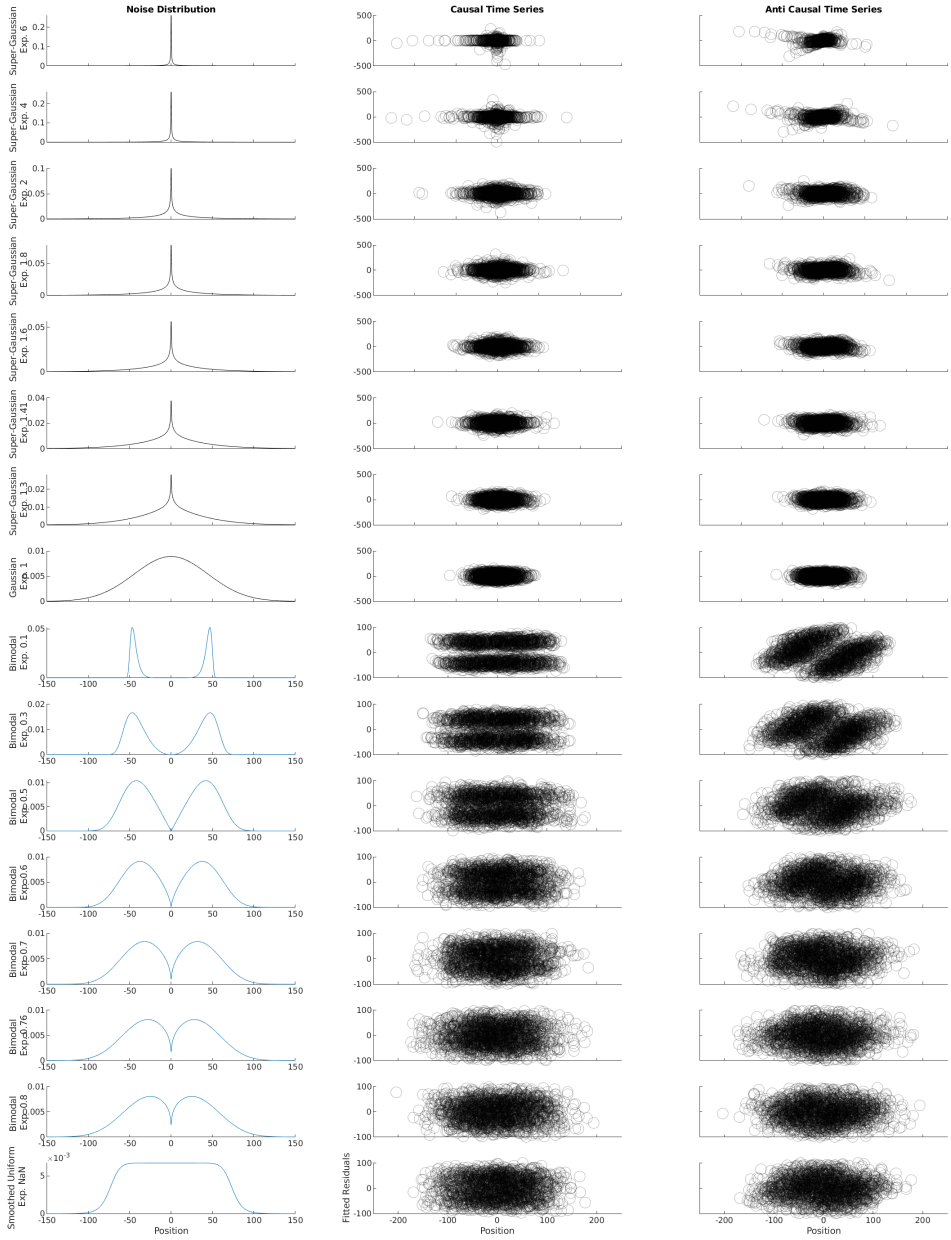

Figure A.1: Noise distributions (left panel) used in our experiments and for generated time series plots of fitted residuals ϵ_t over variables x_{t-1} in causal direction (middle panel) and anti causal direction (right panel). The independence in middle panel and the dependence in the right panel is used by the ResDep algorithm to determine the direction of the time series [37].

A similar approach was chosen for the smoothed Uniform distribution. The smoothed Uniform distribution has the following form

$$p(z|w,c) = \frac{1}{k}\frac{1}{1+(\frac{z^2}{w^2})^c}. \tag{5}$$

The constant k is a normalization term

$$k = \frac{\pi \cdot w}{2 \cdot c \cdot sin(\frac{\pi}{2 \cdot c})}, \tag{6}$$

the term c determines the steepness of the tails and the w variable controls the width of the distribution. We choose to fix the constant c to 6.
One can show that for $c = 6$

$$Var(\epsilon_t) = \frac{1}{k}\frac{w^3\pi}{3 * \sqrt{2}}. \tag{7}$$

We solve this equation again for the width parameter and in the next step for the normalization parameter to equalize the variance to 2000. Values for random noises were generated with the build-in Matlab generator for the Bimodal, Super-Gaussian and Gaussian distribution. Rejection sampling was used for the smoothed Uniform distribution.

A.2 Psychophysical Experiment

A.2.1 Paradigm and procedure

Figure A.2 shows a static stimulus from our experiment (animated stimuli are available in the supplementary material). The position of the moving disk was directly obtained from the time series. A constant was added to centre the time series. The moving disk had a diameter of 1.5 cm, corresponding to 0.6 degrees of visual angle for the 70 cm distance between observers and display, the disk was attenuated with a centered Gaussian distribution of standard deviation of 0.252 cm. Vertical bars show the centre of the screen to help participants in judging the position of the dot. Stimuli were presented against a black background. The colour in normalized values for green dots was (0,1,0), for red dots was (1,0,0) and the bars were again (1,1,1). In response trials, dot colour was white (1,1,1).

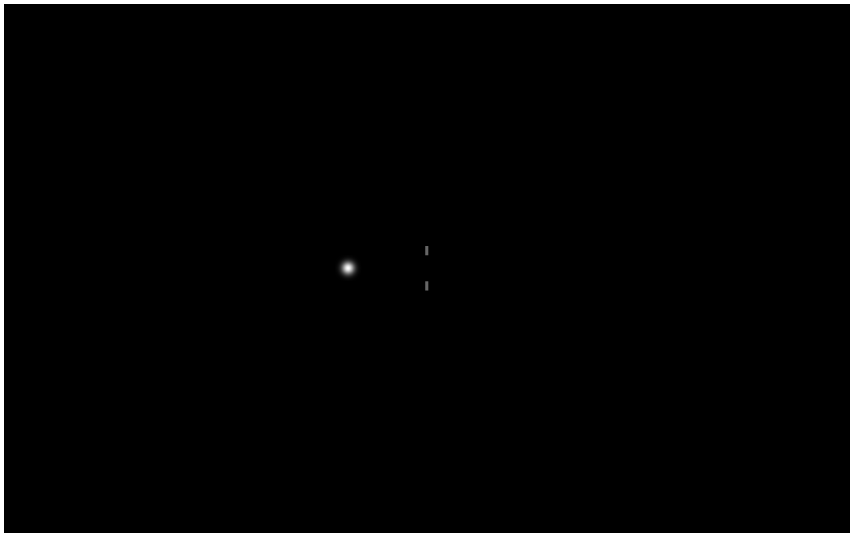

Figure A.2: Static stimuli for the psychophysical experiment. In the learning trials, participants saw the moving dot with the corresponding color. In experimental trials, the color was changed to white and subjects had to press a button for the corresponding color.

Prior to the experiment written consensus was collected from all participants. Observers were told that they should imagine working as doctors and that they view bacteria through a microscope. They have to classify the bacteria in good bacteria (green, causal) and bad bacteria (red, anti-causal). Neither the noise distributions were told to the participants nor any other information about the underlying scope of the experiment.
For each noise distribution (Super-Gaussian, Bimodal, smoothed Uniform, Gaussian) they saw 10 learning trials in the beginning where the dots were already colored. In case of Super-Gaussian and Bimodal, where we could change the difficulty with the exponent, the training trials were the easiest condition (r = 0.1 and r = 6). In training trials dots were already colored and observers could only passively watch the movement, no responses were collected. It usually took participants only a few trials until they reported that they were able to

Table 2: Ordering for stimulus presentation for the first experiment for each subject.

Subject	Day 1		Day 2	
	1.	2.	1.	2.
AW	Su-Gauss.	Gauss	Bimod.	sUnif.
LG	Bimod.	sUnif	Su-Gauss	Gauss
TF	Su-Gauss	Gauss	Bimod.	sUnif.
JS	Bimodal	Gauss	Su-Gauss	sUnif
BW	sUnif.	Bimod.	Su-Gauss.	Gauss
LW	sUnif.	Su-Gauss.	Bimod.	Gauss
JM	Su-Gauss.	Gauss	Bimod.	sUnif.
SM	Bimod.	sUnif	Su-Gauss	Gauss
LL	Su-Gauss	Gauss	Bimod.	sUnif.
PB	Bimodal	Gauss	Su-Gauss	sUnif

discriminate the dots. The ordering of conditions within the Super-Gaussian and Bimodal conditions was the same as in Table 1.

In one session we presented stimuli from the long lasting distributions (Super-Gaussian, Bimodal) and one of the short lasting distributions (Gaussian, smoothed Uniform); we never started with a Gaussian noise distribution, Gaussian noise is not identifiable and we did not want to demotivate observers, see table 2.

One trial consists of a stimulus presentation time of maximum 25 seconds (250ms per position) and an additional response time of 1.2 seconds. The experiment was self-paced, thus participants could indicate their choice with a button press at any time during stimulus presentation; stimulus presentation stopped as soon as a button was pressed. Participants were not allowed to change their choice. Trials with no response were not counted as an answer. After each trial feedback was provided by colouring the dot at the last position either green or red for causal or anti-causal time series. Feedback was provided for one second. Afterwards, we showed a 200ms black screen as inter-stimulus interval.

We calculated the number of trials according to the following rule of thumb. In psychophysical trials the binomial distribution of correct responses is often approximated with a Gaussian distribution. The one-sided 95% confidence interval for a Gaussian distribution with probability p_0 is:

$$p_0 + 1.96 \cdot \sqrt{\frac{p_0(1-p_0)}{N}} \tag{8}$$

We decided that we want to discriminate the chance probability $p_0 = 50\%$ from 60% with a 95% confidence. This yields a sample size of 42 and we finally settled on 40 trials in our experiment per noise distribution and observer.

Every trial was randomly chosen with a probability of 50% to be a causal or an anti-causal trial. For each noise distribution and each exponent, subjects classified 40 Trials in 2 blocks of 20 trials. At the end of each block, the overall accuracy and response time was displayed. The complete experiment lasted on 2 days 1.5 hours each. After finishing the experiment participants received a debriefing.

The paradigm for experiment 2 was slightly chance. We were interested in the effect of shorter viewing times. We used Bimodal noise with exponent $r = 0.5$ since we noticed that participants could classify these time series well. On the other hand, we expected that the accuracy drops fast enough to a range, where subjects showed a performance above change level but below ceiling performance. This helped us us to investigate the relationship between humans and algorithms.
At first, we familiarized subjects again with the task. Subjects observed ten times series with Bimodal noise and exponent $r = 0.1$. Afterwards, Subjects rated 10 time series with increasing difficult Bimodal noise (r = 0.1,0.3 0.5). All participants reported that they were able to classify the time series. Then we fixed the exponent r to 0.5 and reduced the maximum viewing time by reducing the number of time points as indicated in table 3. For a fairer comparison between algorithms and observers, the subjects were explicitly instructed to observe the time series as long as possible unless they were really sure about the trial, see figure A.2.1.

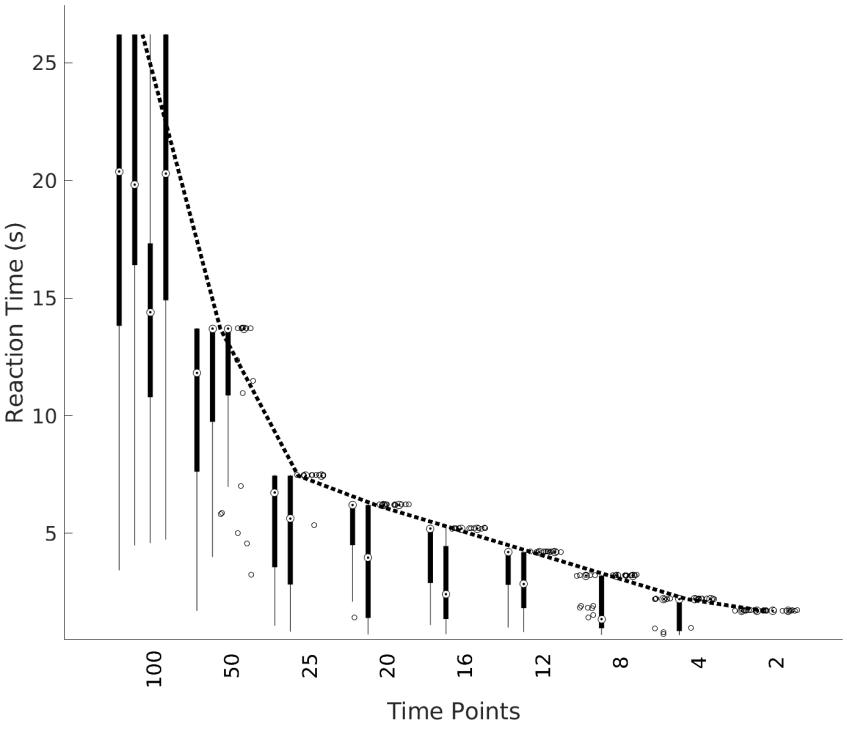

Figure A.3: Boxplots of the reaction times for the four observers in Experiment 2, from left to right subject SB, AG, LL, BW. The dashed line shows the maximum possible reaction time.

Table 3: Ordering for stimulus presentation for second experiment where the viewing time was reduced subsequently. The seed was used to freeze the noise for all subjects. However additionally we checked that all subjects really saw the same time series.

Number of time points	100	50	25	20	16	12	8	4	2
Seed	1001	1002	1003	1004	1005	1006	1007	1008	1009

We encouraged the subject to observe the time series as long as possible to prevent that subjects always answered after a short viewing period. All other parameters were constant and not changed compared to experiment 1. This second experiment lasted 1.5 hours.

A.2.2 Observers

The first experiment was piloted with two naive observers (1 male, 1 female). They showed good performance for all noise distributions. 17 naive observers participated in the first experiment (9 female, 8 male, mean age = 26y stdev= 6.4y). The discrimination task is rather difficult and we screened participants based on their performance in what we considered an "easy conditions" $r = 6, 4, 2$ (Super-Gaussian) and $r = 0.1, 0.3, 0.5$ (Bimodal). This was done at the beginning of the first experimental session when they first proceed to one of the distribution. We excluded 7 of the observers after the first few blocks since they did not get above 67.5% performance which was needed to be significantly different from chance level. This happened in 6 of 7 cases with Super-Gaussian noise. A post-hoc and somewhat fuzzy explanation might be that for Super-Gaussian noise with large exponents only "a cue event" rarely happened. Most of the time noise with small values around 0 is sampled. Thus, it was difficult for them to determine the direction. If they were not concentrating they easily missed a cue event. Thus we think that excluded observers were not in general unable to detect the causal direction (successful observers get easily >90% in these conditions) but that the seven excluded observers were not fully focusing on the experiment. All observers reported normal or corrected-to-normal vision and received monetary compensation and a bonus if they got an accuracy larger than 65% at Super-Gaussian (r = 1.41), bimodal (r=0.76) and smoothed Uniform

noise. The second experiment was performed by the two pilots and 2 good observers from the first experiments (2 male, 2 female, mean age = 23y stdev =2.4y)

A.2.3 Apparatus

Stimuli were presented on a $22''$ VIEWPIxx LCD monitor (VPixx Technologies, Saint-Bruno, Canada) in a dark room. The monitor had a spatial resolution of 1920×1200 pixel (484×302mm) and a refresh rate of 100 Hz. A chinrest and a headrest were used to keep the position of the head constant during the experiment. Response collecting was done with the RESPONSEPixx (VPixx Technologies, Saint-Bruno, Canada) controller. Stimulus presentation and response recordings were controlled using Psychtoolbox 3.0.12 [55, 56] in MATLAB (Release 2016a, The MathWorks, Inc., Natick, Massachusetts, U.S) along with the iShow-library (`https://zenodo.org/record/34217`) on a desktop computer (12 core CPU i7-3930K, AMD HD7970 graphics card "Tahiti" by AMD, Sunnyvale, California, United States) running Debian 9.

A.2.4 Analysis and Algorithms

We describe briefly our way to fit psychometric functions to the data in experiment 1 and experiment 2. Afterwards we describe in depth the 3 algorithms (Dependence, DNN and Bayes) used in comparison to human data.

In general, all analyses are done in MATLAB R2018b, most of the plots were done in R (version 3.4.2). Psychometric functions were fitted with the psychometric toolbox [52] with a fixed guessing rate of 0.5. For experiment 1 a cumulative Gaussian function was used. We fitted a logistic function to the data of Experiment 2. No further changes to the default options were applied.

We used the ResDep algorithm proposed and offered by Bauer et al. [38] to detect the time series based on independence relationships between data and (fitted) residuals. We modified the original algorithm to run in Matlab 2018 since the *vgxset* and *vgxvarx* functions are not supported in the newest MATLAB version. To compare the algorithm to the human performance, we also used a forced-choice paradigm, thus we always force the algorithm to choose the direction in which the residuals are more independent. The algorithm outputs 0 if it fails to fit a time series. We subtracted the mean and divided by the standard deviation each time series as preprocessing step since MATLAB fittings have problems with time series of large variance. We made sure that centering the time series does not change the relationship of residuals to data.
For the second experiment, we slightly further modified the original algorithm. In the original algorithm, the order of the fitted time series is chosen with Akaike Information Criterion [57] between 1 and 10. Instead of the original implementation, we fit only up to order 1 and 5. We changed the maximum fitted order since the Matlab routine has a problem to fit a short time series with long lags. The algorithm starts guessing for time series shorter or equal than 8 time steps because fitting series is not possible and additionally also the independence test does not work anymore.

The neural network was implemented in Matlab with the Deep Learning Toolbox (Version 12.0). Each Network was trained separately for each noise distribution. We generated 30,000-time series, half of it causal one, the other half anti-causal ones. We used 75% as training data and 25% for validation. The first layer consisted of a convolutional layer of kernel size 1×10 and 10 kernels in total, followed by a batch-Normalization layer, a relu-layer, a fully connected Layer and finally a softmax layer for classification. The initial learning rate was set to 0.01 together with the Adam-solver of MATLAB. We limited learning to a maximum of 30 epochs (approximately 3 minutes) and set the "Validation Patience" to 5 epochs. All other values remained to the default values. For every noise, we trained the network 3 times and chose the one with the best performance. This had only very small effects on performance.
For experiment 2 minor modifications were necessary. We reduced the number of time points n until subjects only saw two points in total. Thus we had to change the architecture for the convolutional layer and changed the size of the first convolutional layer to $min(10, n)$.

The last algorithm described is the ideal observer algorithm. We calculated the probability that the time series is in causal direction ($d = 1$) or in anti causal direction ($d = 0$) given the data $X_{t=\{1,...,N\}}$

$$p(d = 1|X) = \frac{p(X_{t=\{1,...,N\}}|d = 1) \cdot p(d = 1)}{p(X_{t=\{1,...,N\}})} = \frac{\Pi_{t=1}^{N} p(x_t|x_{t-1}, ..., x_1) \cdot p(d = 1)}{p(X_{t=\{1,...,N\}})}. \tag{9}$$

since we only use time series of order p = 4 we could limit this to

$$p(d=1|X) = \frac{p(X_{t=1,\ldots,N}|d=1)\cdot p(d=1)}{p(X_{t=\{1,\ldots,N\}})} \tag{10}$$

$$= \frac{\Pi_{t=1}^{N} p(x_t|x_{t-1},x_{t-2},x_{t-3},x_{t-4})\cdot p(d=1)}{p(X_{t=\{1,\ldots,N\}})} \tag{11}$$

$$= \Pi_{t=5}^{N} p(x_t|x_{t-1},x_{t-2},x_{t-3},x_{t-4})\cdot \frac{p(x_4|x_3,x_2,x_1)\cdot p(x_3|x_2,x_1)\cdot p(x_2|x_1)\cdot p(x_1)\cdot p(d=1)}{p(X_{t=\{1,\ldots,N\}})} \tag{12}$$

In the anti-causal direction we get

$$p(d=0|X) = \Pi_{t=1}^{N-4} p(x_t|x_{t+1},x_{t+2},x_{t+3},x_{t+4})\cdot \frac{p(x_{N-3}|x_{N-2},x_{N-1},x_N)\cdot p(x_{N-2}|x_{N-1},x_N)\cdot p(x_{N-1}|x_N)\cdot p(x_N)\cdot p(d=0)}{p(X_{t=\{1,\ldots,N\}})} \tag{13}$$

It follows that under assumption of equal probability of forward and backward time series that

$$\frac{p(d=1|X_t)}{p(d=0|X_t)} \approx \frac{\Pi_{t=5}^{N} p(x_t|x_{t-1},x_{t-2},x_{t-3},x_{t-4})}{\Pi_{t=1}^{N-4} p(x_t|x_{t+1},x_{t+2},x_{t+3},x_{t+4})}. \tag{14}$$

This equation holds only approximately since we skipped the terms after the first expression in eq. (12) and eq. (13). For long time series we omit these terms since they only influence the final ratio marginally. However, this could explain why the Bayes ideal observer algorithms performs for a small number of time points worse than a Neural Network in experiment 2. For short time series the approximation does not hold anymore. The Bayesian ideal observer algorithm always outputs 0 for time series smaller than 5 data points.

A.3 Individual Results for Observers in Experiment 1

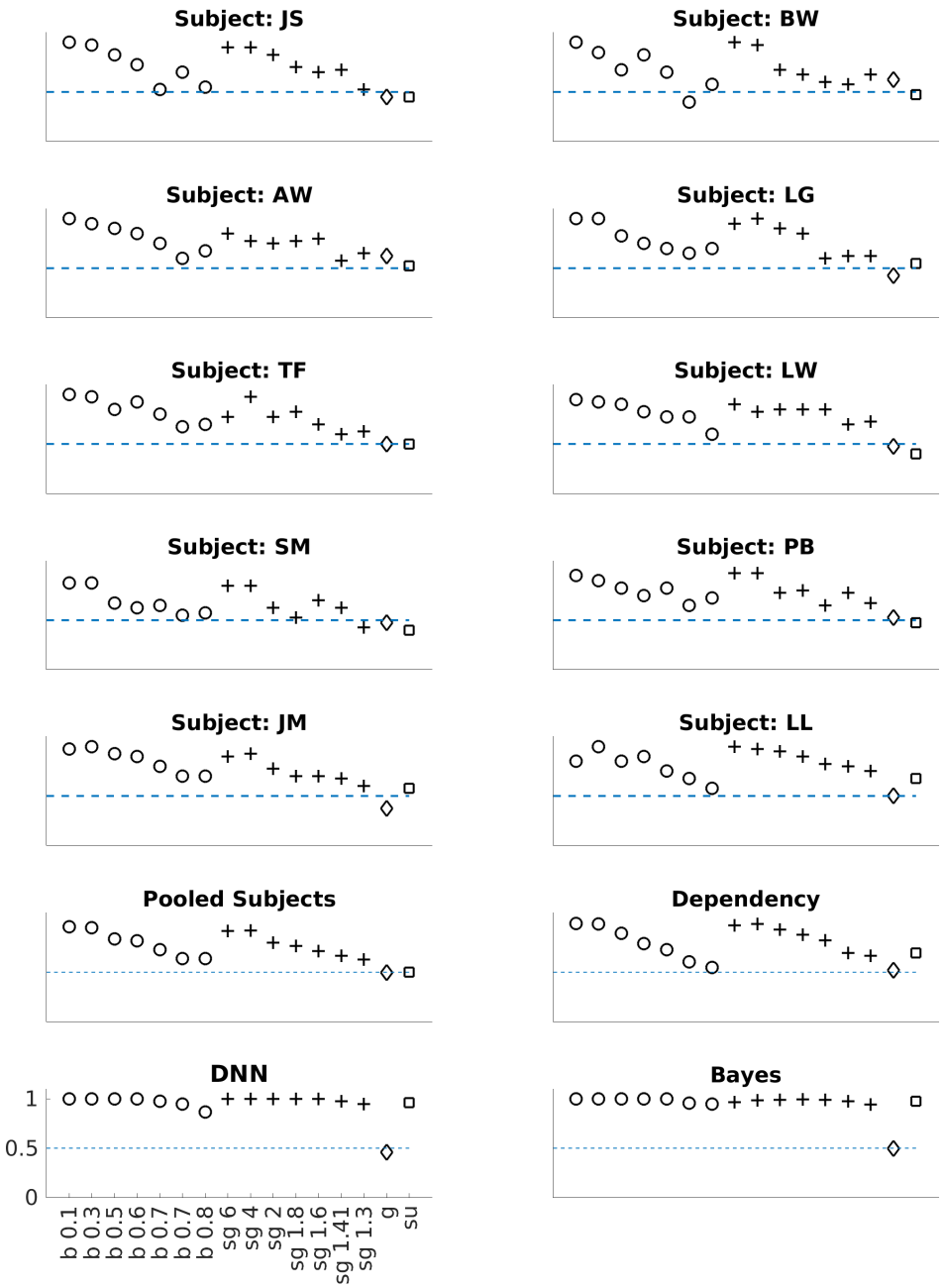

Figure A.4: Accuracy of the ten single observers in experiment 1 across all noise distributions and pooled accuracy for all humans and the algorithms. From left to right we show performance for increasing difficult Bimodal noise(b), Super-Gaussian noise(sg), Gaussian noise(g) and smoothed Uniform(su) noise. The number corresponds to the exponent of the distribution. The horizontal line shows the chance performance of 50%.

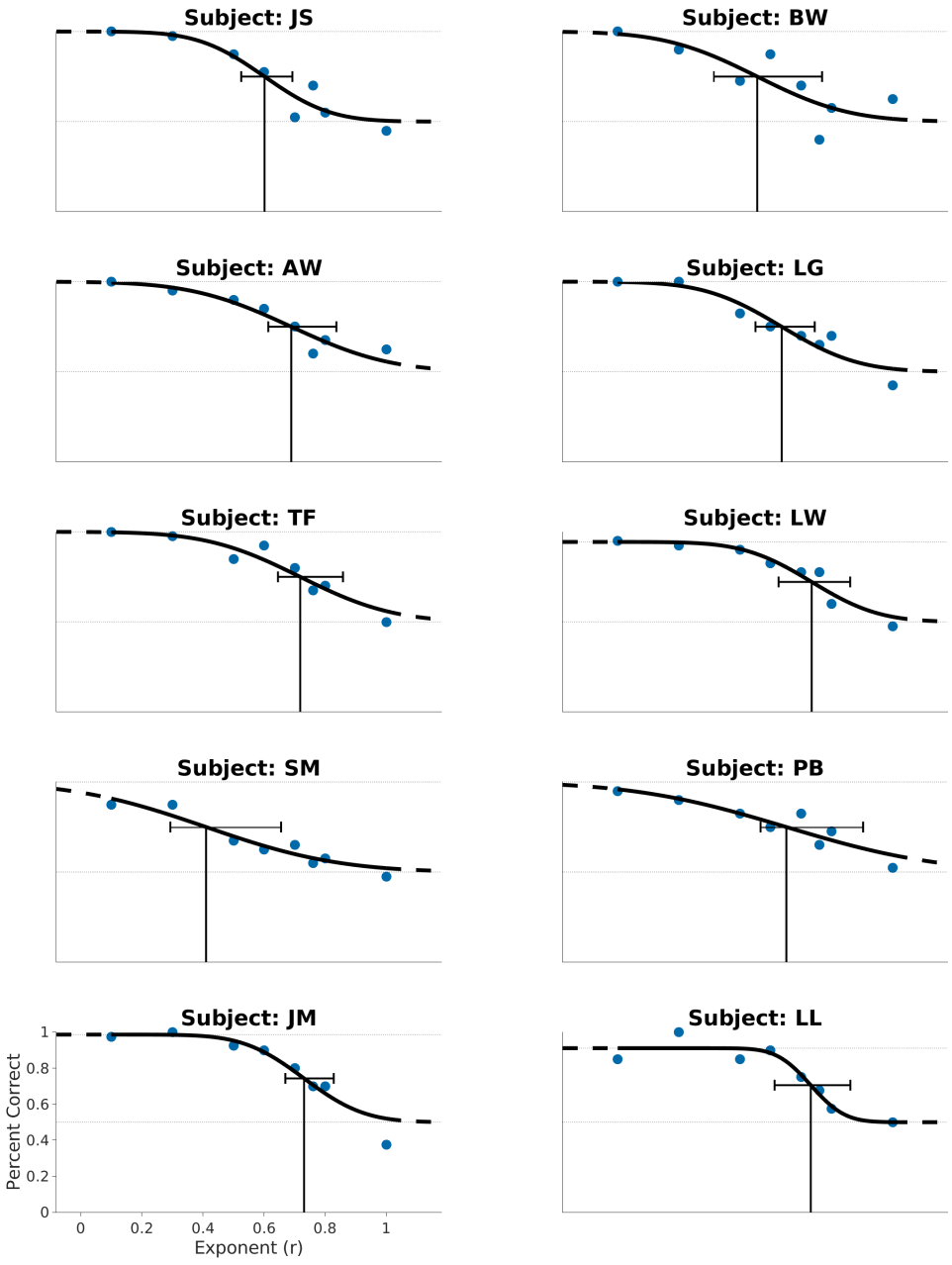

Figure A.5: Psychometric Functions of the ten single observers in experiment 1 for time series with Bimodal noise. Blacks dot represent human mean performance. Vertical lines represent the 75% threshold. Whisker represent the 95% Credible Interval.

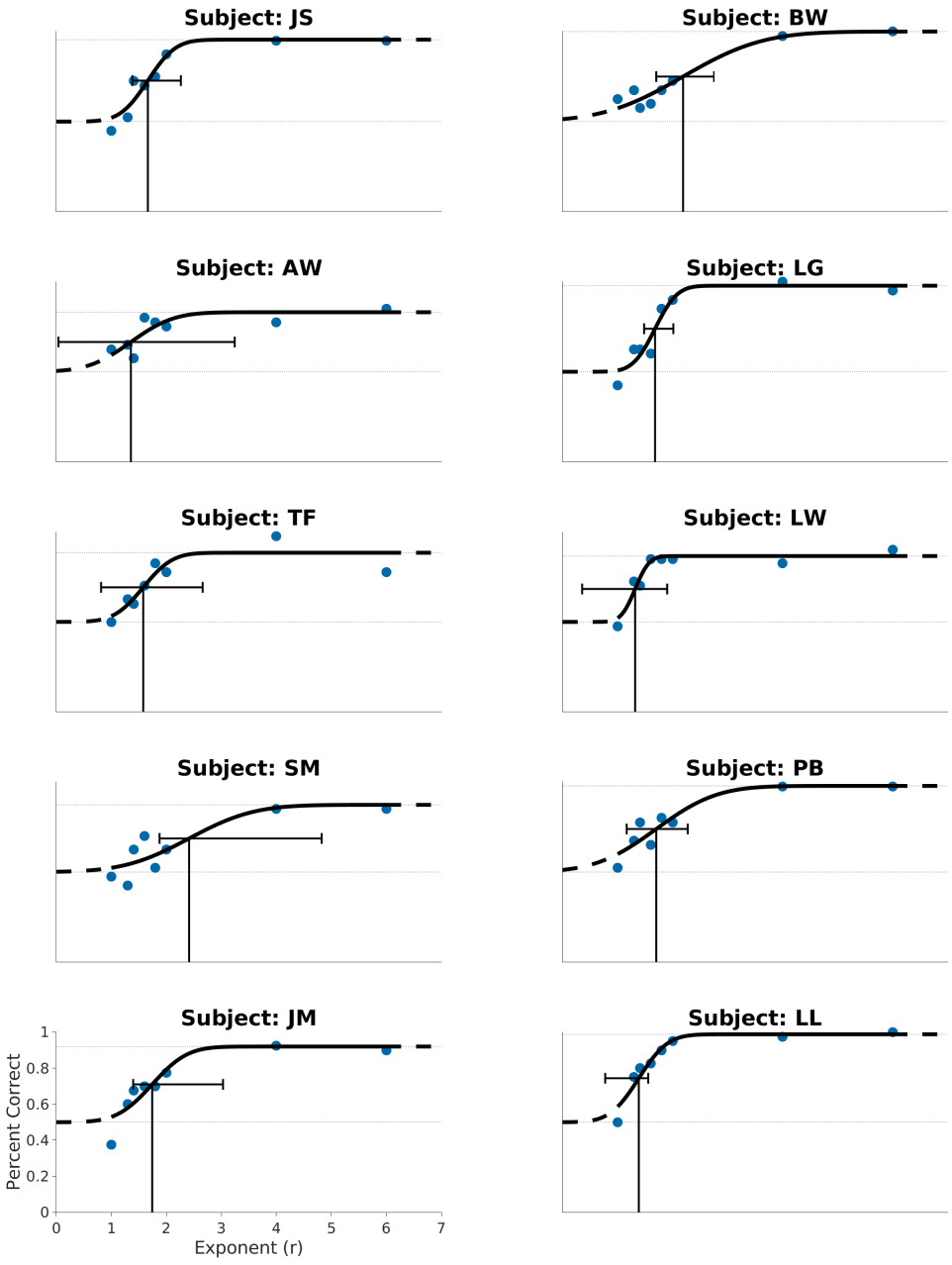

Figure A.6: Psychometric Functions of the ten single observers in experiment 1 for time series with Super-Gaussian noise. Blacks dot represent human mean performance. Vertical lines represent the 75% threshold. Whisker represent the 95% Credible Interval.

Table 4: Thresholds and 95% Credible Interval for the 10 individual observers in experiment 1 from psychometric functions fitted by Psignifit 4[52]

Observer	Threshold Bimodal	Threshold Super-Gaussian
JS	0.60 [0.53,0.69]	1.66 [1.38,2.27]
BW	0.56 [0.41,0.77]	2.19 [1.70,2.74]
AW	0.69 [0.61,0.84]	1.36 [0.04,3.24]
LG	0.64 [0.55,0.74]	1.68 [1.48,2.01]
TF	0.72 [0.65,0.86]	1.58 [0.81,2.66]
LW	0.74 [0.63,0.86]	1.32 [0.35,1.90]
SM	0.41 [0.29,0.66]	2.41 [1.88,4.83]
PB	0.65 [0.57,0.90]	1.70 [1.16,2.27]
JM	0.73 [0.67,0.83]	1.74 [1.40,3.03]
LL	0.73 [0.61,0.86]	1.38 [0.77,1.55]

A.4 Analysis of the neural network

Figure A.4 shows a further performance investigation of the neural network. Learned weights are shown in figure A.8 and figure A.9. We use the trained networks and tested the accuracy on all other noise distributions. The neural network generalizes better from difficult noise (r =0.8 or r=1.3) to easy noise (r=0.1 or r=6) than in the other direction. Furthermore, the network trained on Super-Gaussian noise or Bimodal noise interchange the label if we test for the other noise distribution. This effect is indicated by a performance below chance level and in extreme cases even accuracy of 0%. Bimodal noise and smoothed Uniform noise seem to show similar performance as expected from the residual distribution, see discussion in the previous section. Remarkably, some subjects told us after the experiment that when we switched the noise distributions from day 1 to day 2 they also thought that we interchanged the labels of the time series. Additionally, the network is able to capture the noise distributions when trained on Super-Gaussian and Bimodal noise simultaneously. The Super-Gaussian and Bimodal distribution had same KL-Divergence to a Gaussian with same mean and variance.

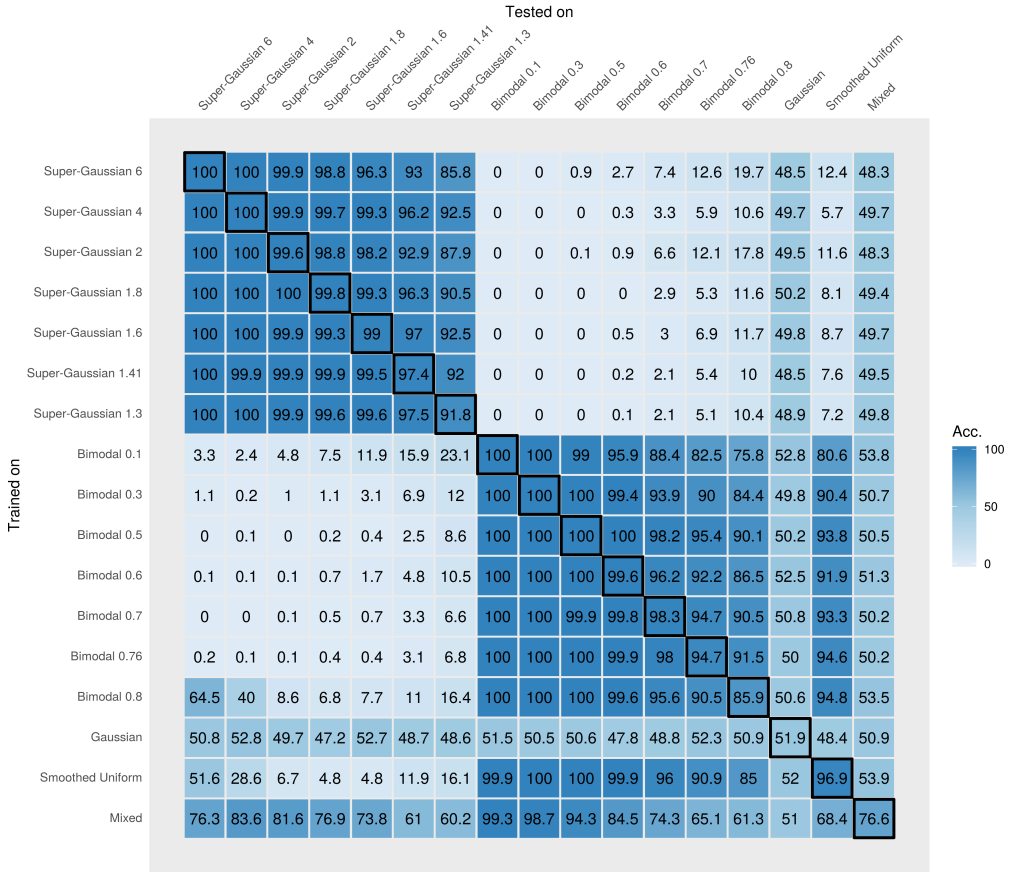

Figure A.7: Generalization of the Neural Network across different noise distributions. We trained the network on one noise distribution(rows) and tested against all other distributions (columns). The mixed condition consists of a dataset where half of the time series have Bimodal noise and the other Super-Gaussian noise with equal KL divergence compared to Gaussian distribution with same mean and variance.

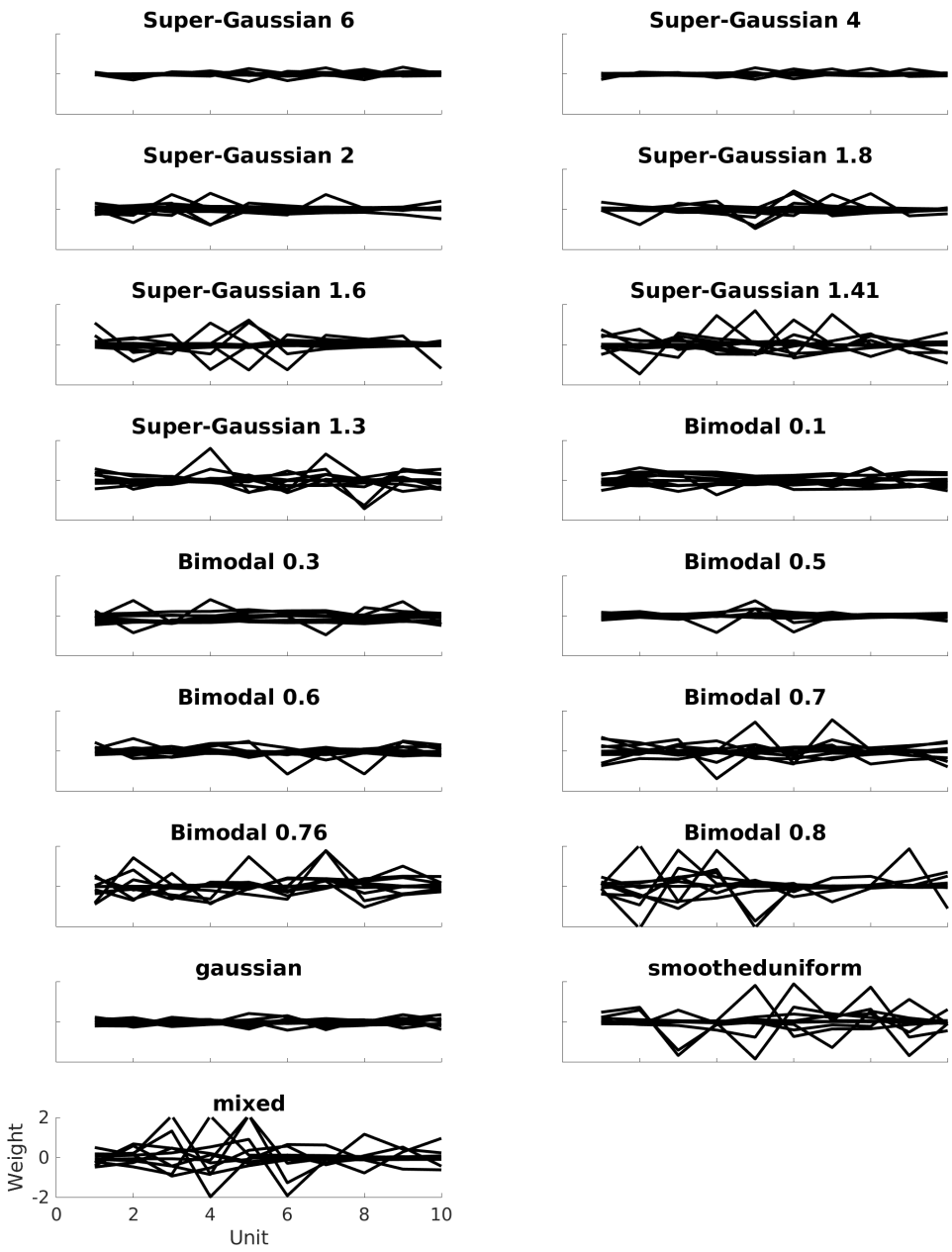

Figure A.8: Weights of the first convolutional layer of the neural network across all noise distributions. We used 10 kernels of length 10. The weights for the easier tasks are (exponents away from 1) are in general smaller than weights for more difficult cases (weights close to 1).

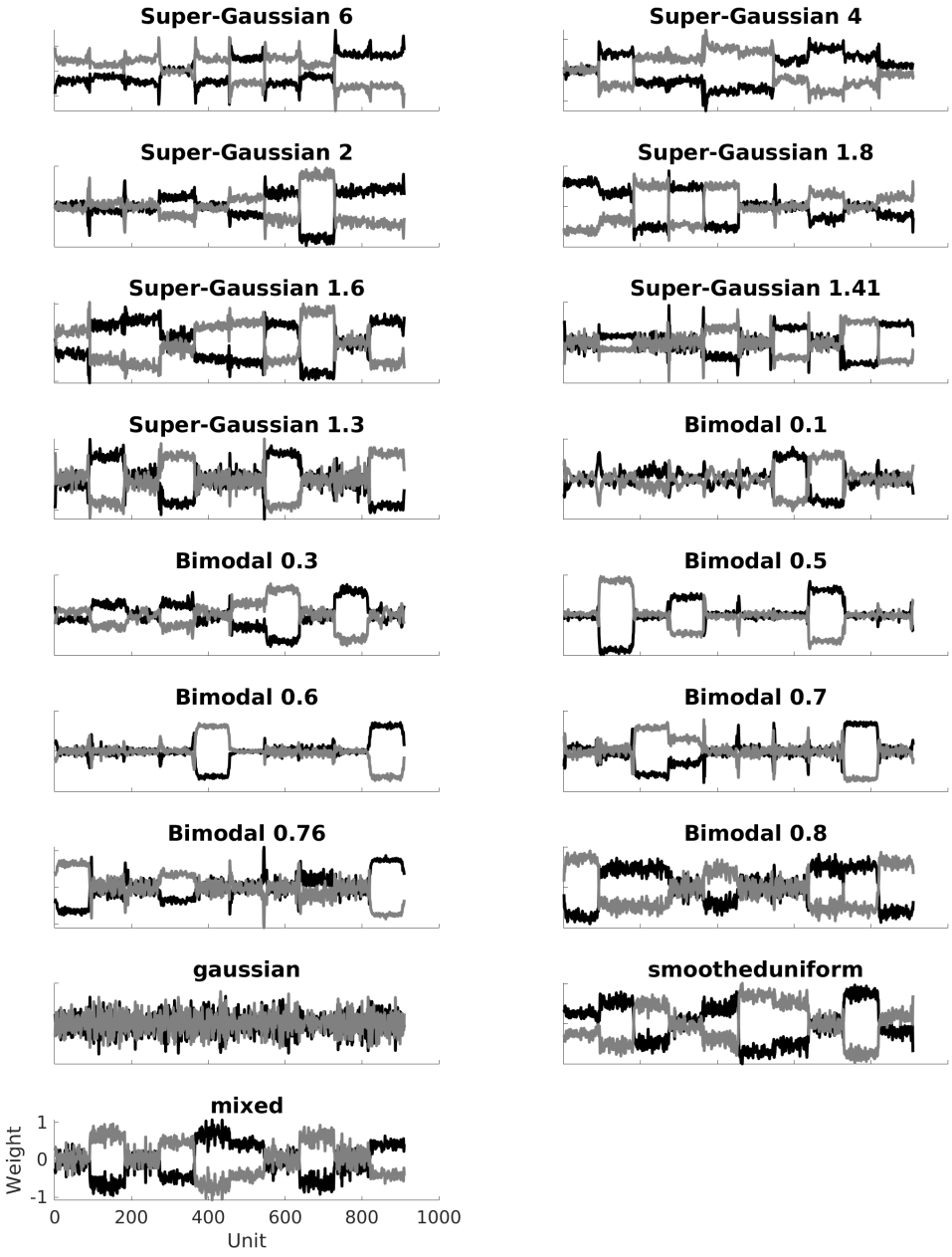

Figure A.9: Weights of the first convolutional layer of the neural network across all noise distributions. Black lines show the weights for output units in causal direction and grey lines the weights for output units in anti-causal direction.

A.5 Performance of human observers in other distance spaces

The parameterization with exponent r is somewhat arbitrary and not directly linked to a parameterized difficulty-scale. A more natural parameterization could be the distance between noise distribution and the non-identifiable Gaussian distribution with the same mean and variance. We calculate the distance with two Information Theory-based f-divergences (Kulback-Leiber Divergence and the symmetric Jensen-Shannon Divergence) as well as the Kolmogorov–Smirnov statistic in figure A.10. Human psychometric functions overlap for the exponent in the JS-space and overlap less in the KL-divergence or KS-distance space. Thus, the performance of humans seems to be captured rather well if we express the distributional distances in the JS-divergence space. Furthermore, if we plot the Bimodal psychometric function on a $1/r$ scale, the difficulty of the Bimodal and Super-Gaussian conditions for human observers is roughly equal, indicating that in our parameterized difficulty for human observers is reasonably well captured by the distance of $1/r$ for the Bimodal and r for the Super-Gaussian noise.

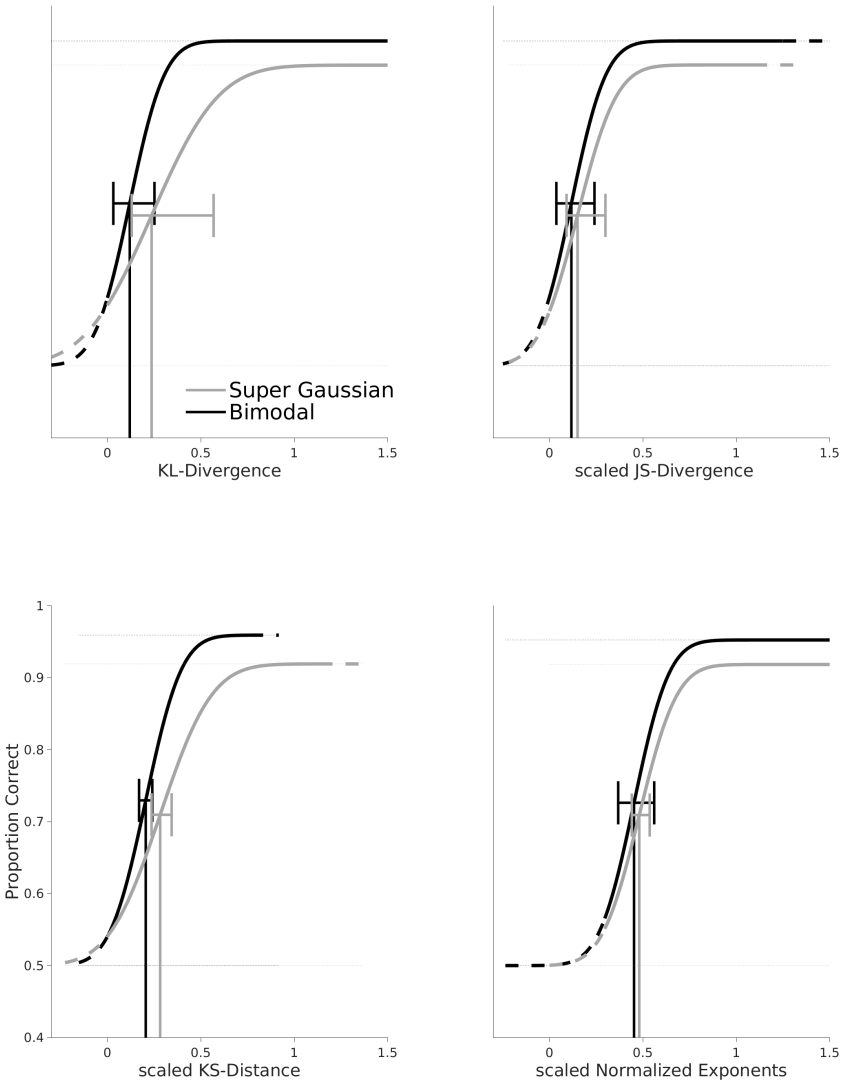

Figure A.10: Psychometric functions of human observers in other distance spaces. The first plot shows the psychometric function where we converted the exponents to a KL-divergence between the noise distribution and a Gaussian distribution with the same mean and variance. The second plot the conversion to the symmetric Jenson-Shanon divergence and the third plot the Kolmogorov-Smirnov distance. The last plot show the psychometric function when we inverted all Bimodal exponents (r<1). We scaled the x-axis for the JS-Divergence, KS-Divergence and Exponents such that all bimodal psychometric functions have the same slope, scale values are 3.5 (JS), 2.91 (KS), 0.296 (Exponents). Thus we can compare the threshold distances graphically. Vertical lines show thresholds and whiskers show the 95% Credible Interval for thresholds from Psignifit.

A.6 Block by block comparison of experiment 1

The psychometric functions summarize performance as a function of a single independent variable. To understand and compare human observers and the different algorithms it is sometimes instructive to compare the performance of humans and algorithms on a block-by-block basis (40 trials per block), as shown in Figure A.11. Each data point represents the performance of one of the 10 observers for one noise distribution on the x-axis plotted against algorithmic performance for the same time series on the y-axis; in addition we compare the performance of the algorithms to each other. The ResDep algorithm is a little better than humans (almost 60% of blocks are above the diagonal). The ideal observer and the neural network are, not surprisingly given figure 1 better for almost every block of 40 trials than the human observers and there is little correlation between human observers and the two algorithms.

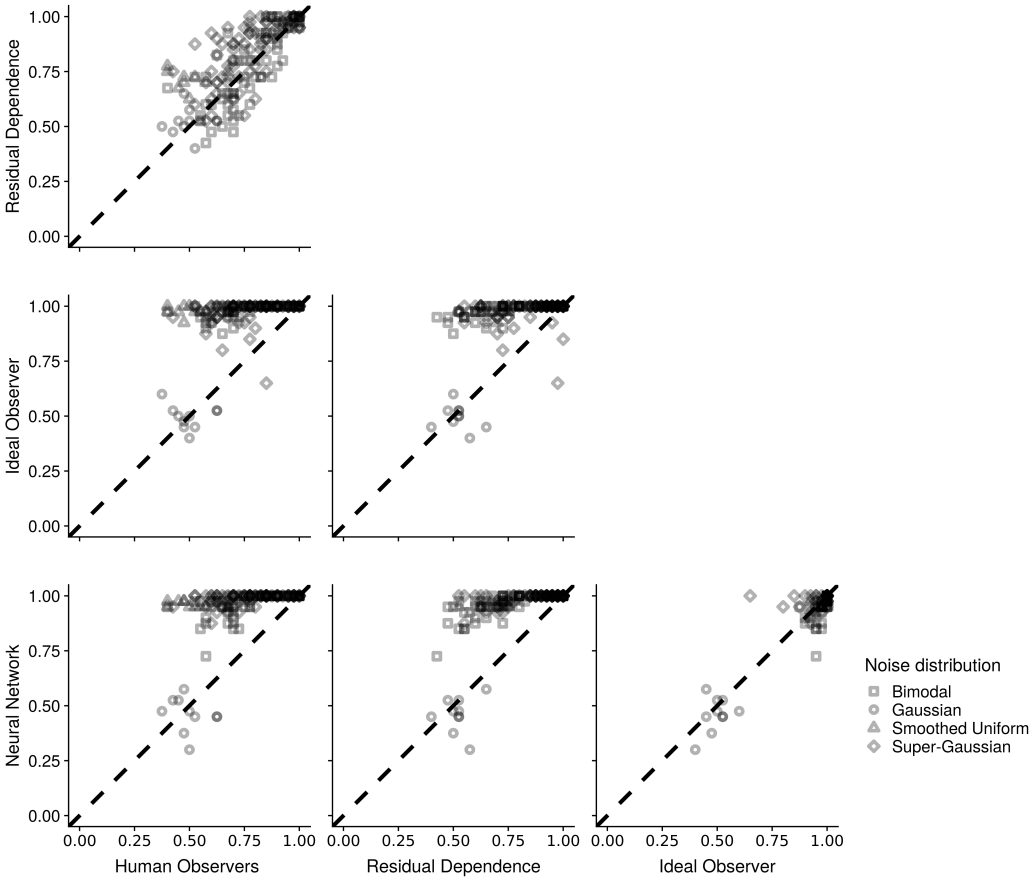

Figure A.11: Performance comparison on a block basis between humans and ResDep, ideal observer and neural network. Every single symbol corresponds to one block with 40 trials. Different symbols correspond to different noise distributions.

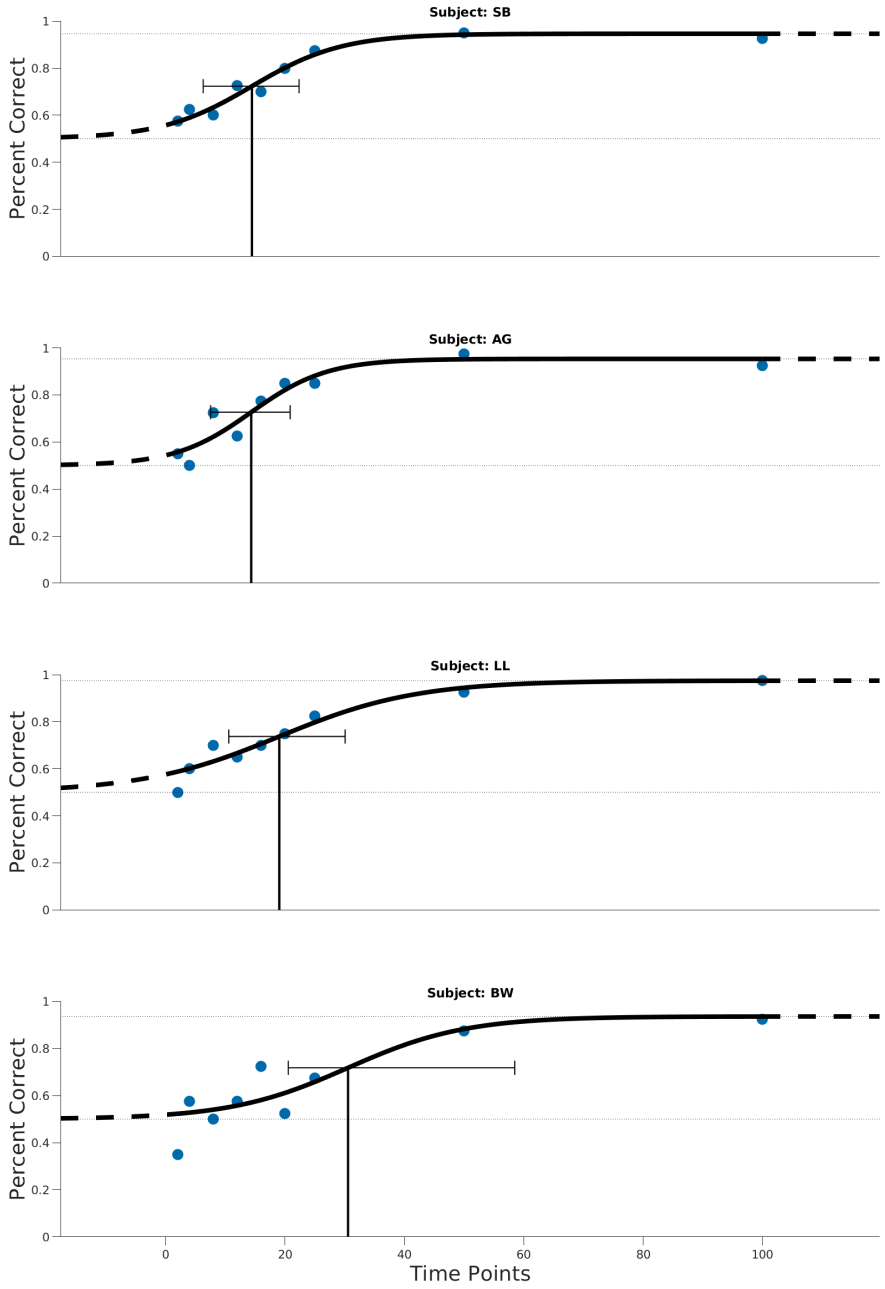

Figure A.12: Individual performance of observers in experiment 4. The vertical line indicates chance performance. All 4 observer show a similar performance even for very short time series. Individual Thresholds are for Subject SB: 14.45, 95% CI [6.29 22.35], Subject AG: 14.33, 95% CI [7.53,20.86], Subject LL: 19.04, 95% CI [10.58,30.07], Subject BW: 30.55.04, 95% CI [20.53,58.48]

A.8 Results of experiment 2 for 3 Observers

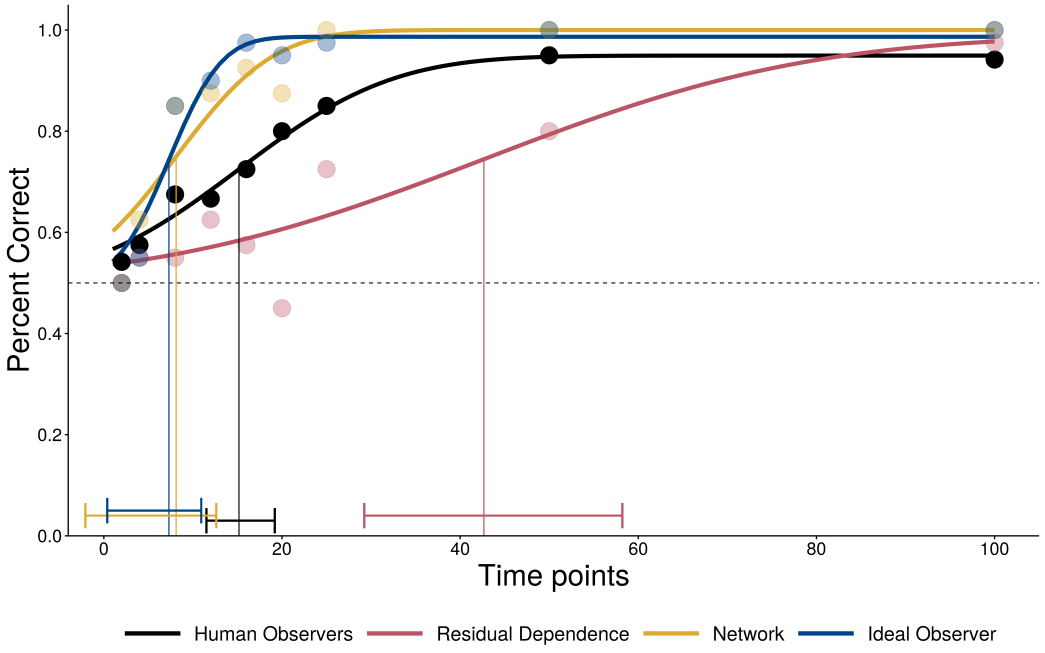

Figure A.13: Individual performance of observers in experiment 4. Similar to figure 2 but excluded poor performing subject BW. Blacks dot represent human mean performance. Vertical lines represent the 75% threshold. Whiskers show 95% confidence intervals of thresholds.

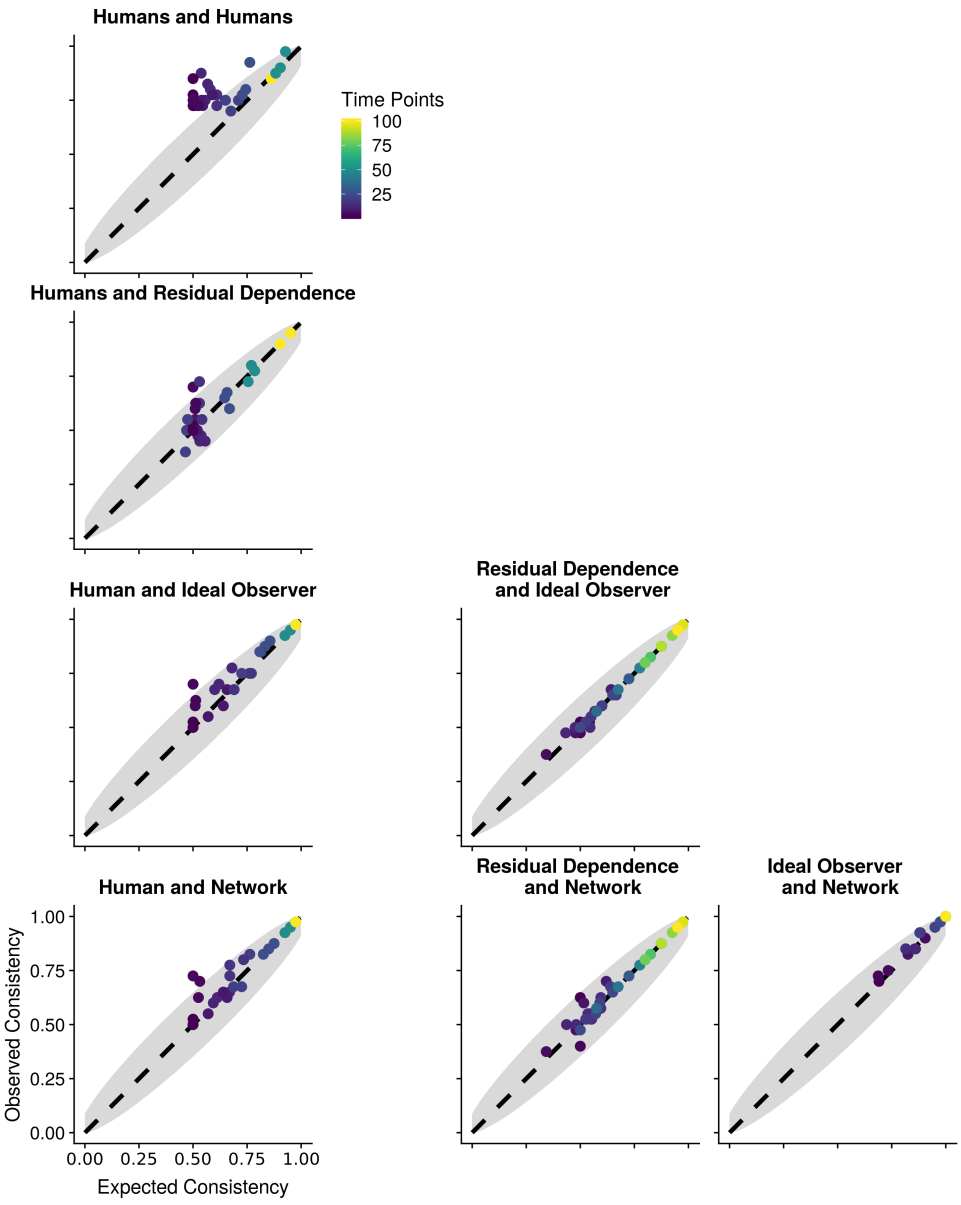

Figure A.14: Human observer consistency and observer-algorithmic consistency for the frozen noise paradigm. Similar to figure 3 but excluded poor performing subject BW. Human observer consistency and observer-algorithmic consistency for the frozen noise paradigm. The x-axis shows the expected proportion of equally answered trials under the assumption of independent observers or algorithms. The y-axis shows the actual observed number of equally answered trials in the experiment. Shaded area shows a 95% confidence interval calculated based on the Wilson score interval [53]. Color codes the number of time points. We used in the algorithm-algorithm comparison not only time series with lengths from the experiment but also a finer grid which: 10-30 time points with spacing 1 and 35-100 with spacing 5.

A.9 Performance of human and algorithms in other distance spaces

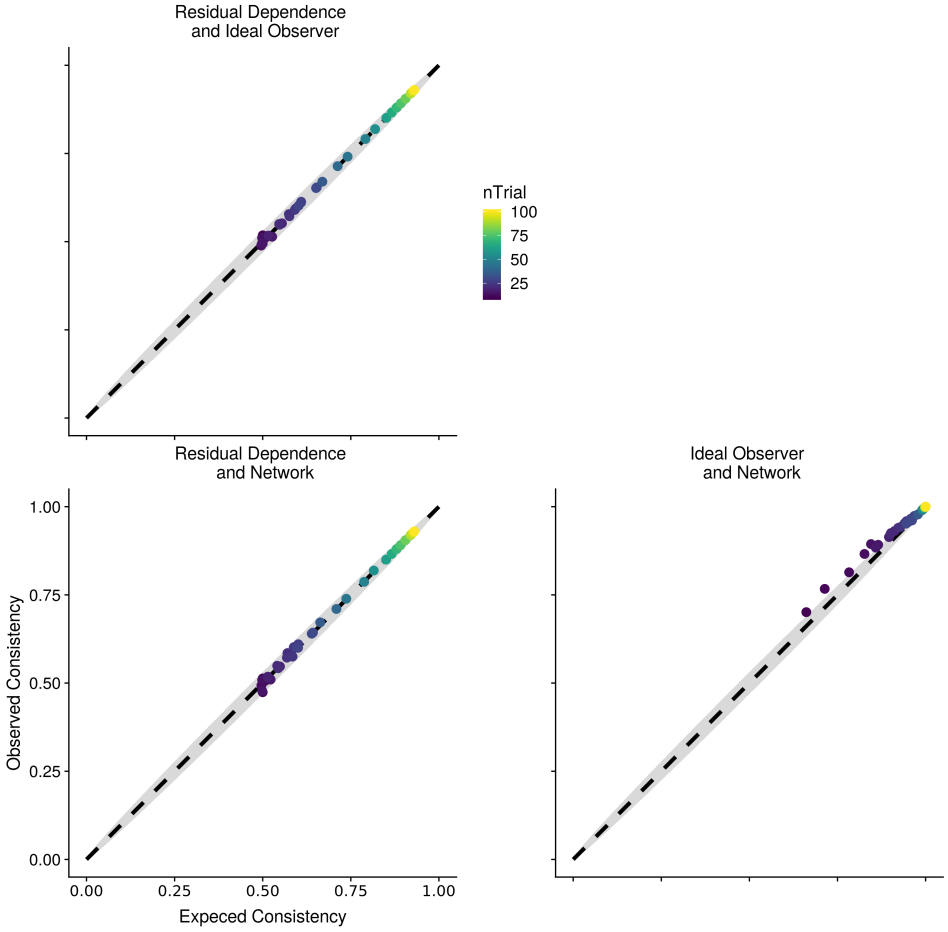

Figure A.15: Revaluation of lower left part in Figure 3 with more trials. Expected consistency versus observed consistency for algorithms for 1000 trials per time point condition. Lengths of time series ranged from 5-31 time point with spacing 1 and 35-100 with spacing 5.

A.10 Frozen noise analysis between ideal observer and network fort difficult Super-Gaussian noise

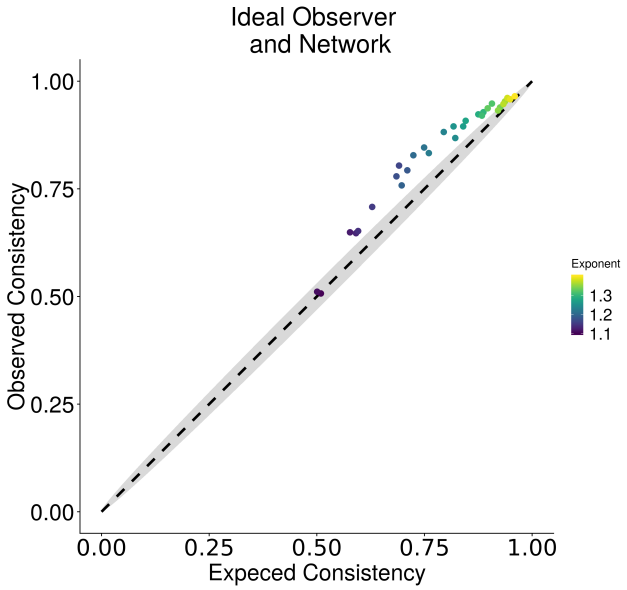

Figure A.16: Ideal observer and network consistency for the frozen noise paradigm. Similar to figure A.15. But we avoid the intrinsic problems of the ideal observer algorithm 2.3 with shot time series by using again time series with 100 time points and made the time series more difficult by making the exponent closer to 1 (Gaussian noise). 1000 times series per exponent were used. The x-axis shows the expected proportion of equally answered trials under the assumption of independent observers or algorithms. The y-axis shows the actual observed number of equally answered trials in the experiment. Shaded area shows a 95% confidence interval calculated based on the Wilson score interval [53]. Color codes the Exponent. We used 30 exponents in the linear range between 1.1 and 1.4.