信頼度推定にもとづく信頼較正キューの選択的提示

Determining Whether to Show Trust Calibration Cues by Inferring Human Reliance

福地庸介 [1*]　山田誠二 [1,2]
Yosuke Fukuchi[1]　Seiji Yamada[1,2]

[1] 国立情報学研究所
[1] National Institute of Informatics
[2] 総合研究大学院大学
[2] SOKENDAI

Abstract: 信頼較正とは、AI システムに何ができて何ができないのか、ユーザに正しく理解させることである。信頼較正のためにシステムがユーザに提示する情報を、較正キューという。本研究では、より少ない較正キューで信頼較正することを目標として、較正キューを提示する必要があるかどうかを判断する手法を提案する。提案手法は、ユーザが AI システムにタスクを任せる確率 (信頼度) を、較正キューを提示する場合としない場合の両条件で予測する。そして、予測された信頼度を AI システムのタスク成功確率と比較することで、キューの信頼較正への影響を評価し、キューを提示するかどうかを判断する。検証の結果、提案手法による較正キューの選択的提示によって、より少ないキューで信頼較正できることがわかった。

1　はじめに

機械学習は、人と協働するコンピュータやロボットを実装する上で有力な手段である。学習モデルの性能向上に伴い、機械学習器を搭載した **AI システム**の応用可能性は広がっている。一方、深層学習器をはじめ、機械学習モデルには人が学習結果を理解することが難しいものがあり、AI システムの適切な利用を難しくしている [1, 2]。

本研究の目的は、AI システムに対するユーザの過信・不信を防止し、また解消する、**信頼較正**である [3, 4]。過信とは、ユーザが AI システムの性能を過剰に見積もっている状態であり、AI システムの誤用やタスクの失敗につながる。逆に不信は、ユーザが AI システムの性能を過小に見積もっている状態で、ユーザは AI システムが活用できる場面でもシステムに頼らなくなるため、ユーザの負荷の増大や、協働のパフォーマンス低下を招く。これまで、AI システムがユーザに対し、較正キューと呼ばれる情報やシグナルを提示することで、ユーザの信頼を調整する試みがなされてきた [5]。

較正キューの提示における課題として、キューをいつ提示すべきかの決定が挙げられる。従来研究の多くでは、画面上にキューを常に提示する状況を想定している。しかし、例えばロボットがキューを口頭で伝達する場合、キューを提示し続けるのはユーザの負担となる。必要なタイミングで必要なキューだけを選択的に提示できれば、人とシステムのコミュニケーションにかかるコストを抑えながら信頼較正できる。

Okamura & Yamada は、ユーザの過信・不信を検知し、それに合わせてキューを提示する手法を提案している [6, 7]。しかしこの手法は、人間がタスクを自分や AI システムに誤って割り当てた回数のみを参照しており、対象となる AI システムとこれまで協働した経験の詳細や、システムがどのようなタスクを行えるかに関するユーザの信念といった、認知的側面は考慮されていない。

本稿では、較正キューを選択的に提示するための手法を提案する。手法の基本となるアイデアは、AI システムに対する人の信頼度と AI システムの信頼性に乖離が生じないように、較正キューを提示するか選択するというものである。ここで、信頼度とは、人間が現在のタスクを AI システムに割り当てる確率のことであり、信頼性は、AI システムがタスクに成功する確率である。提案手法を構成する重要な要素である信頼モデルは、較正キューが提示される場合とされない場合の両方で人の信頼度を予測する。そして、予測された信頼度とシステムの信頼性を比較することで、信頼較正へのキューの寄与度を評価し、キューを提示するかど

*連絡先：国立情報学研究所
〒 101-0003 東京都千代田区一ツ橋 2 丁目 1-2
E-mail: fukuchi@nii.ac.jp

うか判断する。

人と AI システムの協働タスクをクラウドソーシング上で実施し提案手法を検証した結果、較正キューを提示するかランダムに決定した場合は、キューの数が少なくなると協働のパフォーマンスが低下する一方で、提案手法によって決定した場合は、パフォーマンスの低下が見られなかった。この結果は、提案手法が較正キューを提示するかどうかを適切に選択できていることを示唆している。

2 従来研究

2.1 信頼較正

信頼には、直接観測できない心理的態度の側面と、実際に行動として観測できる側面があり、それぞれ trust と reliance として区別される [8]。人と AI システムの協働パフォーマンスを指標とする本研究が着目しているのは reliance であるものの、その関連性の強さから、本章では両者を併せて取り上げる。

信頼較正を行う方法は様々である。Chen et al. が提案する trust-POMDP モデルでは、ロボットがより簡単なタスクから先に取り組むことで人間の信頼を得る、という行動が獲得された [9]。一方、人が AI システムの信頼性を評価するのを助ける情報 (較正キュー) を明示的に提示することで信頼較正する方法もある。較正キューとして一般的なものの一つが、AI システムの決定に関する確信度である [10, 11, 12]。本研究でも、確信度を較正キューとして採用する。

従来研究では、画面上に常に提示される較正キューの有効性が示されてきた [12, 13]。しかし、較正キューを常に提示することには少なくとも 2 つの懸念がある。1 つは、伝達する方法によってはコミュニケーションのコストが大きくなることである。例として、ロボットが口頭で較正キューを提供する場合が挙げられる。もう 1 つの懸念は、常に提示される較正キューに対して人が注意を向けなくなることである。Okamura & Yamada の実験では、AI システムの信頼度を提示し続けたにもかかわらず、ユーザの過信が修正されない例や、AI システムがトリガーとなる刺激を追加で与えることが有効であることが報告されている [6]。そこで本研究は、較正キューを選択的に提示することで信頼較正することを目指す。

較正キューの選択的提示に関連する工学的手法はほとんどない。Okamura & Yamada の提案する手法は、人の過信と不信を数学的に表現した「信頼方程式」にもとづいて、較正キューを提示するか判断する [6, 7]。しかし、この手法は人が AI システムや自分自身にタスクを割り当てた回数のみを考慮しており、人がシステムの失敗をどのタスクで観察したか、システムはいつキューを提供したか、その時のタスクは何だったかといった、それまでの協働の経験を詳細に捉えることができないという課題がある。このような経験は、AI システムの能力に関する人の信念に影響を与える可能性がある。例えば、あるタスクでシステムが成功／失敗した経験は、異なるタスクよりも類似のタスクにおける人の信頼度に、より影響を与える可能性が高い。提案手法では、人と AI システムの協働の履歴をもとに人の信頼度を予測する信頼モデルを学習しており、こうした側面を捉えることが期待される。

2.2 信頼度の推定

本稿で提案する手法の基本的なアイデアは、AI システムに対する人間の信頼度の推定が、較正キューの選択的提示に寄与するというものである。例えば、人の信頼度を高める較正キューは、人が既にシステムに高い信頼度を持つ場合には、信頼度が低い場合よりも効果が少なくなると考えられる。

信頼度の推定には、質問紙による信頼尺度が一般的に用いられている [14, 15, 16]。また、fMRI や EEG を用いて信頼を推定する研究もある [17, 18]. これらの手法の弱点は、タスクの実行を阻害する可能性がある点である。本研究に関連するアプローチとして、人の行動に着目したものがある。Walker らは、視線の動きから人間の信頼を推測する方法を提案した [19]. また、ロボットの行動に対する人間の介入や乗っ取りは、信頼度が低いという指標になる [20]。人がタスクを自分に割り当てるか AI システムに割り当てるか、という決定も重要な要素であり [6, 7]、本研究でもこれに着目する。

信頼度を推定する手法は多く提案されているものの、較正キューが信頼度に与える影響を考慮したものはない。すなわち、これまでにどのような場面でどの較正キューが提供されたか、現在のタスクに較正キューが提供されるかどうかが信頼度に与える影響を考慮することはできない。

3 較正キューの選択的提示

3.1 定式化

本稿では、較正キューを伴う人と AI システムの協働を、要素の組 $(x, \hat{c}, c, d, y^*, y, p)$ で定式化する。人と AI システムがタスク $\{x_i\}_{i=1}^N$ を逐次的に行う状況を考える。i はタスクの順番を、N はタスクの総数を示す。$\hat{c}_i$ は、x_i が与えられた時に潜在的な較正キューであり、人に提示するかを提案手法が決める。c_i は、実際に人に与えられる較正キューである: 較正キューが提示され

た場合は $c_i = \hat{c_i}$、提示されない場合は $c_i = [MASK]$ と表すことにする。

人は (x_i, c_i) を観測し、x_i を自身か AI システムに割り当てる ($d_i \in \{\text{AI}, \text{human}\}$)。$y_i^*$ は x_i に期待される結果、d_i による実際の結果を y_i とする。x_i に成功した際は $y_i^* = y_i$ となる。人は $d_i = \text{AI}$ の時のみ AI システムの結果を観測できる。p_i は AI システムが x_i に成功する確率である。

3.2 提案手法

提案手法の基本的なアイデアは、AI に対する人の信頼度 r_i と p_i の乖離を避けるように、較正キューを提示するか決めるというものである。信頼モデルは、二種類の r_i を予測する:

$$\begin{aligned} r_i^{\text{w/}} &= P(d_i = \text{AI} | x_{:i}, c_{:i-1}, c_i = \hat{c_i}, d_{:i-1}, y_{:i}^*, y_{:i-1}), \\ r_i^{\text{w/o}} &= P(d_i = \text{AI} | x_{:i}, c_{:i-1}, c_i = [MASK], \\ &\qquad d_{:i-1}, y_{:i}^*, y_{:i-1}). \end{aligned}$$

$r^{\text{w/}}$ と $r^{\text{w/o}}$ はそれぞれ、$\hat{c_i}$ が提示された時とされていない時の信頼度である。添字 $*_{:i}$ は、系列 $(*_1, *_2, .., *_i)$ を示す。Δ_i は r_i と p_i の乖離の大きさを示す。

$$\begin{cases} \Delta_i^{\text{w/}} &= |r_i^{w/} - p_i|, \\ \Delta_i^{\text{w/o}} &= |r_i^{w/o} - p_i|. \end{cases}$$

$\Delta_i^{\text{w/}}$ と $\Delta_i^{\text{w/o}}$ を比較することで、$\hat{c_i}$ を提示するか決定する:

$$c_i = \begin{cases} \hat{c_i} & (\Delta_i^{\text{w/o}} - \Delta_i^{\text{w/}} < threshold) \\ [MASK] & (\text{elsewise}). \end{cases}$$

$threshold$ は、$\Delta_i^{\text{w/}}$ を基準としたときの $\Delta_i^{\text{w/o}}$ の許容幅を表す。$threshold$ が 0 の場合、$\hat{c_i}$ を提示しない方が乖離を小さくできると予測される時のみ較正キューを省略することになる。また、$threshold$ を大きくすると、より多くのキューが省略されるようになる。

3.3 信頼モデル

信頼モデルは $r_i^{\text{w/}}$ と $r_i^{\text{w/o}}$ を予測するモデルである。モデルは、Transformer エンコーダ [21] をもとに実装した。

信頼モデルは、協働の履歴 $(x_{:i-1}, c_{:i-1}, d_{:i-1}, f_{:i-1})$ と現在のタスク (x_i, c_i) を入力として与えられる。履歴には、いつ、どのタスクにおいて、どの較正キューが提示されたか、人が、自身とシステムのどちらをタスクに割り当てたか、といった情報を含む。これにより、協働の中で人が形成する、AI システムの得意・不得意に関する信念を捉え、システムに対する人の信頼度をより正確に予測することが期待される。

モデルに入力されるそれぞれの要素は、パーセプトロンによって埋め込みベクトルに変換される。埋め込みベクトルには、インデックス情報を付与する position embedding が足し込まれる [21]。さらに、Transformer encoder モデルを経たインデックス i のベクトルが、多層パーセプトロンによって r_i に変換される。

人が AI システムの信頼性を考える上で、システムの結果を y^* と比較するのは重要である。しかし、本研究では AI システムが完璧ではない状況を想定しているため、信頼モデルにも y^* は与えない事とし、代わりとして、人間からのフィードバックに当たる情報 f を与える事とした:

$$f_i = \begin{cases} 0 & (d_i = \text{AI}) \\ 1 & (d_i = \text{human かつ } y_i\text{が AI の結果と同じ}) \\ 2 & (d_i = \text{human かつ } y_i\text{が AI の結果と異なる}). \end{cases}$$

i 番目のタスクにおいて、d_i と f_i は r_i を予測する段階で得られない情報であるため、$[MASK]$ とした。

信頼モデルは教師あり学習で訓練できる。訓練には、二値交差エントロピー誤差関数を用いた:

$$L = -\delta(d_i, \text{AI}) \cdot \log(r_i) - \delta(d_i, \text{human}) \cdot \log(1 - r_i),$$

ここで $\delta(a, b)$ は、$a = b$ の時に 1 を、それ以外の時に 0 を返す関数である。

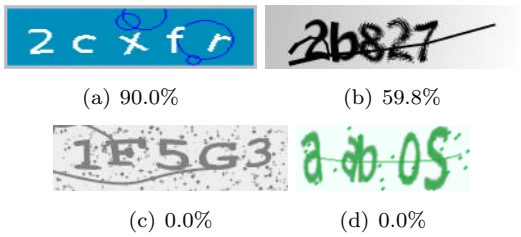
信頼モデルは、$c_t = \hat{c_t}$ と $c_t = [MASK]$ の両方の場合を予測する。その結果がそれぞれ $r^{\text{w/}}$、$r^{\text{w/o}}$ である。

4 信頼モデルの訓練

4.1 協働 CAPTCHA タスク

信頼度モデルの訓練と提案手法の評価のため、タスクとして協働 CAPTCHA を開発した。CAPTCHA は、読みにくく加工された画像に書かれた文字を人が入力するタスクである [22]。協働 CAPTCHA では、参加者が AI システムと協働して文字を入力する。

協働 CAPTCHA において、x_i は CAPTCHA の画像、y_i^* は正答、$\hat{c_i}$ は x_i に対する AI システムの回答の確信度とする。参加者はまず、画像を見て d_i を選択する。$d_i = \text{AI}$ の場合、AI システムが画像を認識し、回答をテキストボックスに入力し、回答送信ボタンが表示される。参加者は回答を送信する前に AI システムの回答を見て、それが正しいか判断することができるが、回答を編集することはできない。$d_i = \text{human}$ の場合、テキストボックスが表示され、参加者が自身で回答を入力する。$N = 60$ とした。



(a) 90.0% (b) 59.8%

(c) 0.0% (d) 0.0%

図 1: CAPTCHA データセットの例。上 2 つを訓練データセットとして用いた。数字は、各データセットに対する AI システムの精度。

4.2 タスクの実装

4.2.1 AI システムの訓練データセット

図 1 に、検証で用いた CAPTCHA 画像の例を示す。4 つのデータセットのうち 2 つのみを、AI システムの訓練データセットとした。これは、データセットの偏りが AI システムの性能に与えるバイアスを再現するためである。参加者が性能のバイアスを理解できていれば、より少ない較正キューでも適切にタスクを割り当てられる。図 1 では、各データセットに対する AI システムの精度も示しており、実際に精度が訓練データセットのバイアスを受けていることがわかる。

それぞれの CAPTCHA には 5 文字が書かれている。AI システムは、j 番目の文字 $x_{i,j}$ が $\iota \in I$ である確率の分布を出力する。ここで I は、アルファベットと数字の集合である。

$$\mathrm{TaskAI}(x_{i,j}, \iota) = P(x_{i,j} = \iota).$$

$d_i = \mathrm{AI}$ の時、 y_i は $x_{i,j}$ の確率が最も高い $\iota \in I$ の系列である。

$$y_i^{\mathrm{AI}} = \{\mathrm{argmax}_\iota(\mathrm{TaskAI}(x_{i,j}, \iota))\}_{j=1}^5.$$

AI システムは、画像認識で一般的なモデルの一つである ResNet-18 を用いた。

4.2.2 AI システムの確信度と較正キュー

確信度は、前節の確率分布から計算した [23]:

$$\hat{c}_i \propto \Pi_{j=1}^5(\max_{\iota \in I}(\mathrm{TaskAI}(x_{i,j}, \iota))).$$

最も可能性が高いと AI システムが計算する文字の確率が高いほど、$\hat{c}_i$ は大きくなる。AI システムの信頼性 p_i は、$\hat{c}_i$ をロジスティック回帰モデルの入力とすることで計算した。このモデルは、訓練データセットにおいて y_i^{AI} が y_i^* と一致するか予測するよう訓練した。

予備実験において、確信度を使った方が信頼較正の性能が良かったため、本研究では信頼性ではなく確信度を $\hat{c}_i$ として使用した。これは、確信度の値が平坦に分布するのに対し、ロジスティック回帰の出力する信頼性の値は 0%と 100%の付近に集中することが影響している可能性がある。

4.3 信頼データの収集と信頼度モデルの訓練

信頼度モデルを訓練するために、協働 CAPTCHA における人の判断 (信頼データ) を収集した。信頼データは、組 $(x, \hat{c}, c, d, y^*, y, p)$ で記述される。

信頼データは、Yahoo! Japan クラウドソーシングを用いて収集した。250 人の参加者が、100 円の謝金で収集された。実験は WEB ベースで行った。参加者には、まず実験に関わる説明が提示され、全員が参加に同意した。続いて、協働 CAPTCHA に関する説明と、タスクの理解を問う質問を 5 問出題した。質問に正答できなかった 99 人はタスクから除外し、151 人が残った (女性 60 人、男性 91 人; $20 - 78$ 歳, $M = 47.5, SD = 12.2$)

x_i はテスト用のデータセットから参加者毎にランダムに割り当てた。ただし、AI システムに対する過剰な過信や不信を生じさせないため、タスク全体での AI システムのタスク正答率が 50%になるように調整した。信頼データの収集では、参加者に較正キューを提示するかはタスク毎にランダムに決定した。較正キューを提示する回数は、$0, 20, 40, 60, 80, 100\%$ となるように統制した。

獲得された信頼データをもとに、信頼度モデルを訓練した。k-fold 交差検証を行った結果、信頼度モデルが d_i を予測する精度は 81.6% (95% CI: 80.0%, 83.2%) だった。

5 評価

5.1 目的

提案手法による較正キューの選択的提示によって、信頼較正できるか検証した。具体的には、提案手法によって、較正キューの数を減らしても人の過信・不信を防ぐことができるか調べた。

5.2 方法

協働 CAPTCHA をタスクとして採用した。参加者は信頼データの取得と同じ手順で AI システムと協働した。ただし、ここでは較正キューの提示タイミングがランダムではなく、提案システムの決定を用いた。91 人の参加者はいずれも、信頼データの取得には参加

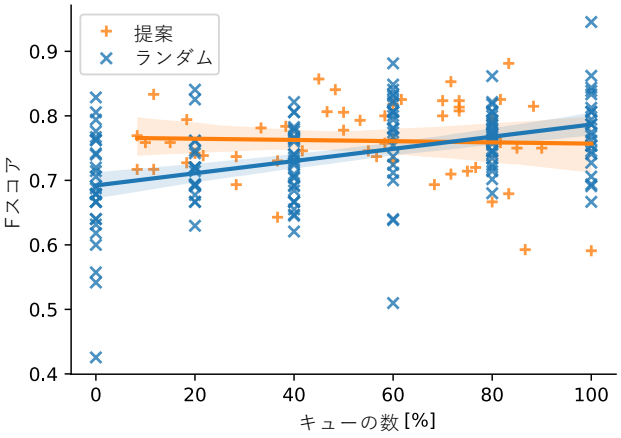


図 2: F スコア。エラーバーは線形回帰の 95%信頼区間。

していない。タスクの理解度を問う質問に正答できた 52 人のみが、協働 CAPTCHA タスクに参加した (女性 14 人, 男性 38 人;21-65 歳, M=43.3, SD=10.2)。タスク実施後には、自由記述でタスクの感想を問うアンケートも行った。*threshold* は、信頼データから計算される $\Delta_i^{\mathrm{w/o}} - \Delta_i^{\mathrm{w/}}$ の分布をもとに、較正キューの数が $20, 40, 60, 80\%$ になるよう設定した。

評価の指標として、人のタスク割り当てと AI システムのタスク成功が一致する度合いを F スコアで計算した。この値を、信頼データから計算される、較正キューがランダムに提示される場合 (ランダム条件) と比較した。

5.3 仮説

提案手法が較正キューを提示するか適切に選択することで、較正キューの数を減らしても、F スコアの落ち込みをランダム条件より抑えられる。

5.4 結果

図 2 に結果の F スコアを示す。結果を統計的に分析するために共分散分析を行ったところ、較正キューが提示された回数 ($F(1, 199) = 30.1; p < .0001; \eta_p^2 = .132$)、キューの提示方法 ($F(1, 199) = 4.54; p = .034; \eta_p^2 = .022$)、交互作用 ($F(1, 199) = 7.31; p = .007; \eta_p^2 = .035$) に有意な差が見られた

ランダム条件では、較正キューの数が減るのに従って F スコアが減少した。提案手法では、得られたデータの範囲内ではキューの数が減ってもスコアの減少が抑えられた。そのため、キューの数が少なくなるほどランダム条件と提案手法条件の間で F スコアの差が拡大している。この結果は、提案手法によって較正キューを減らしてもスコアの減少が抑えられていることを示唆しており、仮説を支持している。以上のことから、人の信頼度予測をもとに較正キューを評価し、必要なキューを選択的に提示することで、提案手法より少ない較正キューで信頼較正を行えていると結論づけられる。

6 おわりに

本研究の重要な限界の 1 つは、人がタスクに失敗する可能性を提案システムが考慮していないことである。2 人の参加者は、自身がタスクに正答する自信がなかったときに AI に割り当てたと回答した。実際に、協働タスクにおける参加者の精度は 84.4%で完璧ではなかった。　本研究では、参加者が AI システムの成功・失敗に対応してタスク割り当てを決定することができたか、という観点で有効性を確認したものの、協働における全体のパフォーマンスを最大化させるためには、参加者のタスク成功確率も考慮する必要があるだろう。

本稿では、人と AI システムの協働において、較正キューを選択的に提示するための手法を提案した。提案手法は、タスクの成功確率と、ユーザの AI に対する信頼度が一定以上乖離しないように、較正キューを提示するか決定する。提案手法の重要な要素として、人の信頼度を、較正キューが提示された場合とされなかった場合の両方を予測する信頼度モデルがある。提案手法を協働 CAPTCHA タスクに実装し、検証を行った。結果、提案手法によって較正キューを選択的に提示する場合、較正キューの数を削減しても信頼較正を達成できており、確信度の予測をもとに較正キューの信頼較正への影響を評価することで、較正キューを提示するか適切に決定できることがわかった。

謝辞

本研究の一部は、JST CREST (JPMJCR21D4) の支援を受けた。

参考文献

[1] Amina Adadi and Mohammed Berrada. Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, 6:52138–52160, 2018.

[2] Arun Rai. Explainable ai: From black box to glass box. *Journal of the Academy of Marketing Science*, 48(1):137–141, 2020.

[3] Kevin Anthony Hoff and Masooda Bashir. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, 57(3):407–434, May 2015.

[4] Raja Parasuraman and Victor Riley. Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, 39(2):230–253, 1997.

[5] Ewart J. de Visser, Marvin Cohen, Amos Freedy, and Raja Parasuraman. A design methodology for trust cue calibration in cognitive agents. In Randall Shumaker and Stephanie Lackey, editors, *Virtual, Augmented and Mixed Reality. Designing and Developing Virtual and Augmented Environments*, pages 251–262, Cham, 2014. Springer International Publishing.

[6] Kazuo Okamura and Seiji Yamada. Adaptive trust calibration for human-ai collaboration. *PLOS ONE*, 15(2):1–20, 02 2020.

[7] Kazuo Okamura and Seiji Yamada. Empirical evaluations of framework for adaptive trust calibration in human-ai cooperation. *IEEE Access*, 8:220335–220351, 2020.

[8] Nicolas Scharowski, Sebastian A. C. Perrig, Nick von Felten, and Florian Brühlmann. Trust and reliance in xai – distinguishing between attitudinal and behavioral measures. In *CHI 2022 Workshop on Trust and Reliance in AI-Human Teams*, 2022.

[9] Min Chen, Stefanos Nikolaidis, Harold Soh, David Hsu, and Siddhartha Srinivasa. Trust-aware decision making for human-robot collaboration: Model learning and planning. *J. Hum.-Robot Interact.*, 9(2), jan 2020.

[10] John McGuirl and Nadine Sarter. Supporting trust calibration and the effective use of decision aids by presenting dynamic system confidence information. *Human factors*, 48:656–65, 02 2006.

[11] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, pages 295–305, New York, NY, USA, 2020. Association for Computing Machinery.

[12] Tove Helldin, Göran Falkman, Maria Riveiro, and Staffan Davidsson. Presenting system uncertainty in automotive uis for supporting trust calibration in autonomous driving. In *Proceedings of the 5th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, AutomotiveUI '13, pages 210–217, New York, NY, USA, 2013. Association for Computing Machinery.

[13] Renate Häuslschmid, Max von Bülow, Bastian Pfleging, and Andreas Butz. Supportingtrust in autonomous driving. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces*, IUI '17, pages 319–329, New York, NY, USA, 2017. Association for Computing Machinery.

[14] Jiun-Yin Jian, Ann M. Bisantz, and Colin G. Drury. Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics*, 4(1):53–71, 2000.

[15] Maria Madsen and Shirley Gregor. Measuring human-computer trust. In *Proceedings of the 11 th Australasian Conference on Information Systems*, pages 6–8, 2000.

[16] Rosemarie E. Yagoda and Douglas J. Gillan. You want me to trust a robot? the development of a human-robot interaction trust scale. *International Journal of Social Robotics*, 4:235–248, 2012.

[17] Anne-Kathrin J. Fett, Paula M. Gromann, Vincent Giampietro, Sukhi S. Shergill, and Lydia Krabbendam. Default distrust? An fMRI investigation of the neural development of trust and cooperation. *Social Cognitive and Affective Neuroscience*, 9(4):395–402, 11 2012.

[18] Sanghyun Choo and Chang S. Nam. Detecting human trust calibration in automation: A convolutional neural network approach. *IEEE Transactions on Human-Machine Systems*, 52(4):774–783, 2022.

[19] Francesco Walker, Willem Verwey, and Marieke Martens. Gaze behaviour as a measure of trust in automated vehicles. 06 2018.

[20] Moritz Krber, Eva Baseler, and Klaus Bengler. Introduction matters: Manipulating trust in automation and reliance in automated driving. *Applied Ergonomics*, 66:18–31, 01 2018.

[21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[22] Luis von Ahn, Manuel Blum, Nicholas J. Hopper, and John Langford. Captcha: Using hard ai problems for security. In Eli Biham, editor, *Advances in Cryptology — EUROCRYPT 2003*, pages 294–311, Berlin, Heidelberg, 2003. Springer Berlin Heidelberg.

[23] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pages 1321–1330. JMLR.org, 2017.