Information Quality Talk Series

-

Valentin Guigon

University of Maryland

Managing Information Under Uncertainty

ISC
Marc Jeannerod

cnrs

GATE
Lyon / St-Etienne

Social Learning & Decisions
SLD Lab

DEPARTMENT OF
PSYCHOLOGY

ARTIFICIAL INTELLIGENCE
INTERDISCIPLINARY INSTITUTE
AT MARYLAND

PROGRAM IN
NEUROSCIENCE &
COGNITIVE SCIENCE

Thinking
About Thinking

ISC Marc Jeannerod

cnrs

GATE Lyon / St-Etienne

Social Learning & Decision SLD Lab

DEPARTMENT OF PSYCHOLOGY

ARTIFICIAL INTELLIGENCE INTERDISCIPLINARY INSTITUTE AT MARYLAND

PROGRAM IN NEUROSCIENCE & COGNITIVE SCIENCE

Thinking About Thinking

Metacognition biases information seeking in assessing ambiguous news

Valentin Guigon, Marie Claire Villeval & Jean-Claude Dreher

Communications Psychology **2**, Article number: 122 (2024) | Cite this article

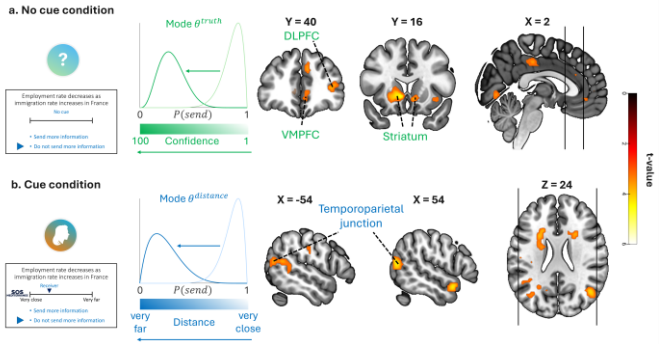
Rethinking misinformation through plausibility estimation and confidence calibration

Valentin Guigon, Lucille Geay & Caroline J. Charpentier

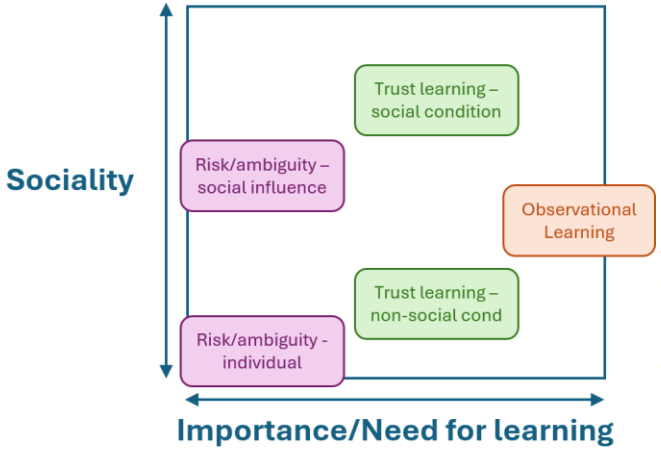
Communications Psychology **4**, Article number: 24 (2026) | Cite this article

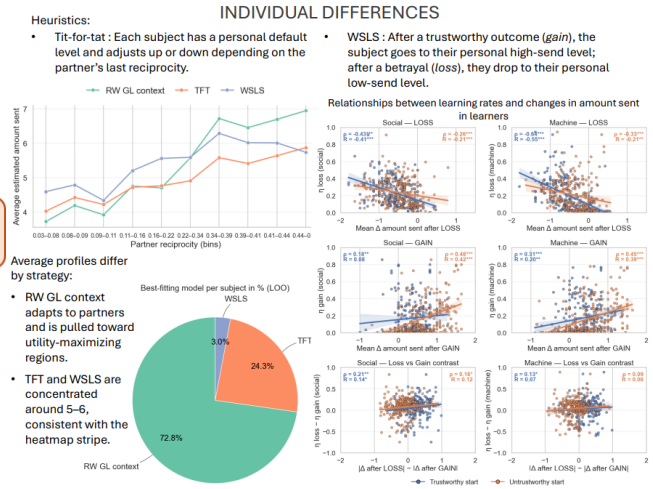
Neurocomputational mechanisms of sharing clarifying information about ambiguous news

V. Guigon[a,b], J. Benistant[a], M. C. Villeval[b] and J-C. Dreher[a]



- **Neurocomputational mechanisms of social learning**
- **Cognitive heterogeneity in solving trust learning games**
- **Computational phenotyping of social learning**





Integrating theory-driven and data-driven computational psychiatry

Samantha Mombelli[1], Samaneh Roghani Zanjani[1], Kian Godhwani[1], Alban Voppel[1,2], Aryamman Aiyaar[3], Tihare Zamorano[1], Timothy Friesen[1], Valentin Guigon[4], Caroline J. Charpentier[4,5], Franziska Knolle[6], Ryan Smith[7,8], Jerome Waldispuhl[9], Nace Mikus[10], Christoph Mathys[10], Andreea O. Diaconescu[11,12,13,14], James Gilleen[15], Lena Palaniyappan[1,2,3], Jai Shah[1,2], Albert Powers[16], David Benrimoh[1,2]

valentinguigon.com

Rethinking misinformation through plausibility estimation and confidence calibration

Valentin Guigon, Lucille Geay & Caroline J. Charpentier

Communications Psychology **4**, Article number: 24 (2026) | Cite this article

> These interventions often assume that improving individuals' ability to detect errors will lead them to revise their judgments and ultimately form more accurate beliefs. This frames misinformation as a problem of truth detection. While normatively appealing, this is difficult to reconcile with how people actually process information—under uncertainty, limited resources, and with multiple goals.
>
> Thus, their ability to identify true from false claims depends less on the quality of their logical reasoning than on how well their confidence tracks a claim's plausibility given limited time, knowledge, and attention.

Better judgment of information requires better information environments and **better strategy selection and calibration**.

AI is now fully part of our information environments and can decide for us. Does our diagnosis still stand?

Transformations of human judgments in the age of AI

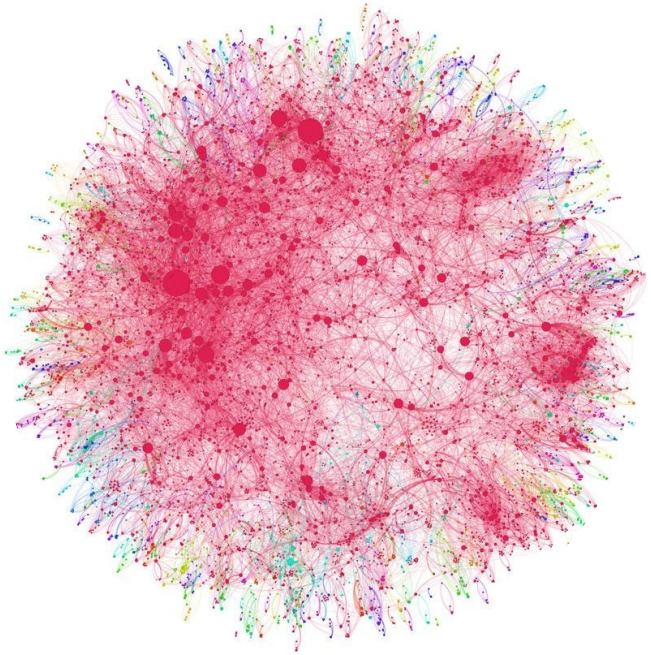
I. Complexities of our information environments

II. How do process and make sense of low-level information?

III. How do social and/or mediated information complexify judgments?

IV. How are human judgments expected to transform with AI?

Complexities of our information environments



Worlds of known & unknown unknowns

Attributed to Leonard Savage (in The Foudation of Statistics, 1954) by Gigerenzer, 2025. *Behavioural Public Policy.*

Table 1. Risk, ambiguity, uncertainty, and intractability

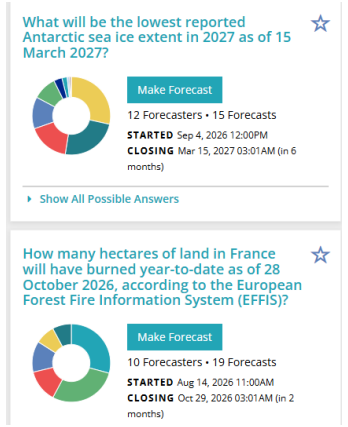
	Small worlds		Large worlds	
Conditions	Risk	Ambiguity	Uncertainty	Intractability
Are all possible future states and consequences of all actions known?	Yes	Yes	No	Yes
Are all probabilities known?	Yes	No	No	Yes
Can optimal action be calculated?	Yes	Yes	No	No

We mostly live in large worlds

	Small worlds		Large worlds	
Conditions	Risk	Ambiguity	Uncertainty	Intractability
Are all possible future states and consequences of all actions known?	Yes	Yes	No	Yes
Are all probabilities known?	Yes	No	No	Yes
Can optimal action be calculated?	Yes	Yes	No	No

Easy problems

Hard problems



Messy real-world problem but exploitable data structure

Messy real-world problem and hardly exploitable regularities

Largeness and resources constraint problem-solving

	Small worlds		Large worlds	
Conditions	Risk	Ambiguity	Uncertainty	Intractability
Are all possible future states and consequences of all actions known?	Yes	Yes	No	Yes
Are all probabilities known?	Yes	No	No	Yes
Can optimal action be calculated?	Yes	Yes	No	No

Easy problems

Hard problems

Messy real-world problem but exploitable data structure

Messy real-world problem and hardly exploitable regularities

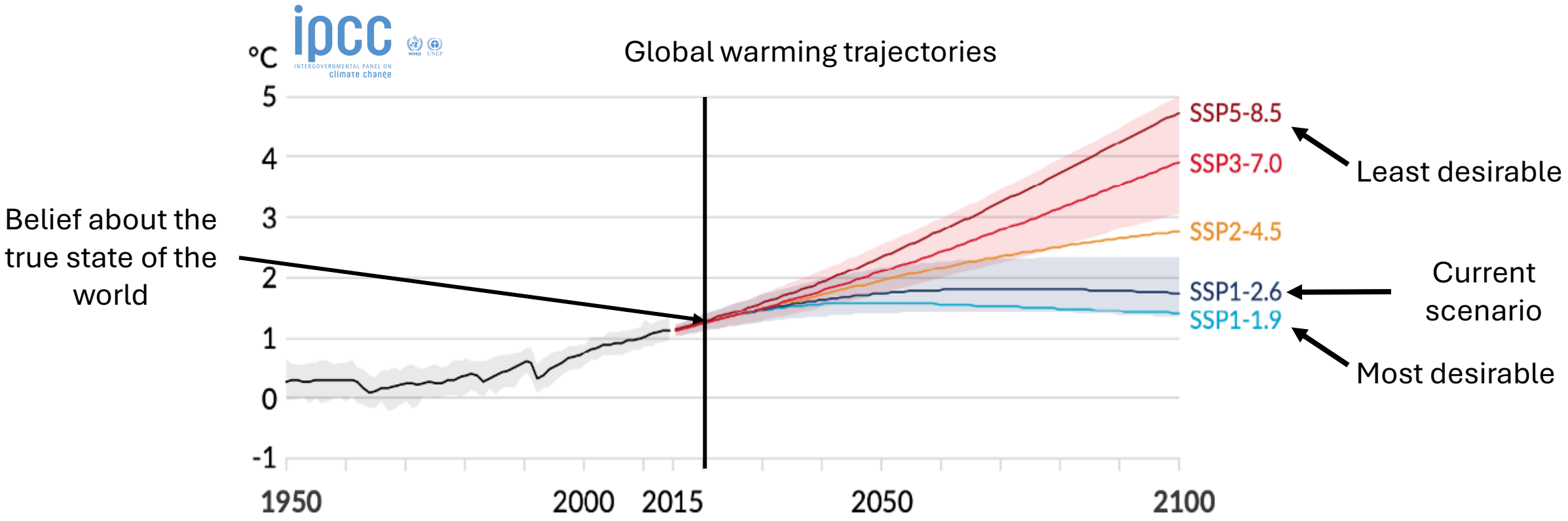
The superiority of a problem-solving strategy depends on the environment's structure exploitability:

- **The structure and predictability of the environment** (largeness, noise, relationships cues-outcomes)
- **The quantity and quality of available information** (richness, relevance, redundancy)
- **Resources** (number of relationships observed, time and computational resources)
- …
  - **Ability to control variance** (bias–variance trade-off, overfitting)

- **Small worlds:** Optimization is feasible and normatively appropriate
- **Large worlds**: Pure optimization is infeasible; simple strategies (heuristics) are more adapted especially for easy problems or hard problems under time and computation constraints

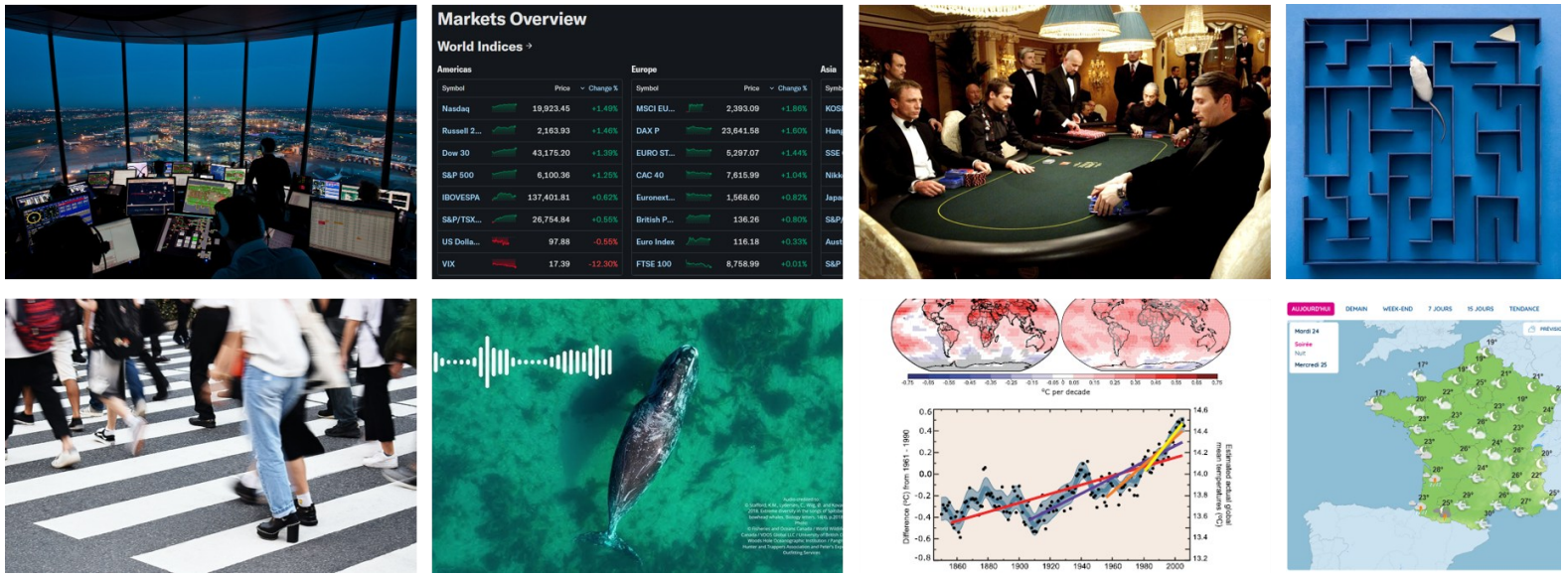
Problem selection -> information-seeking -> estimating true states of the world -> action

Friston et al., 2015. *Cog. Neuro.*; Charpentier et al., 2018. *PNAS*; Shalvi et al., 2019



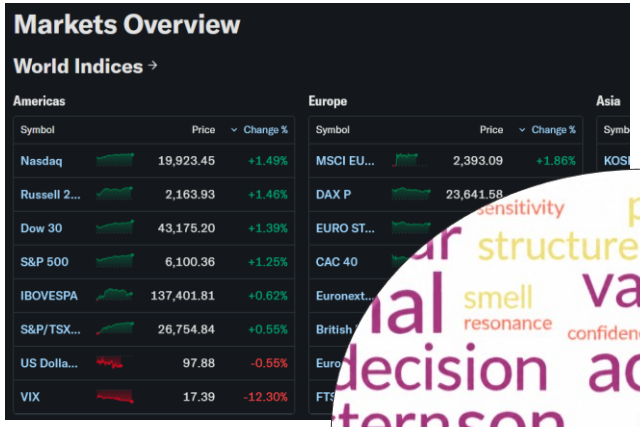
Information value is tied to its **capacity to decrease uncertainties** on contingencies about the world (epistemic value)

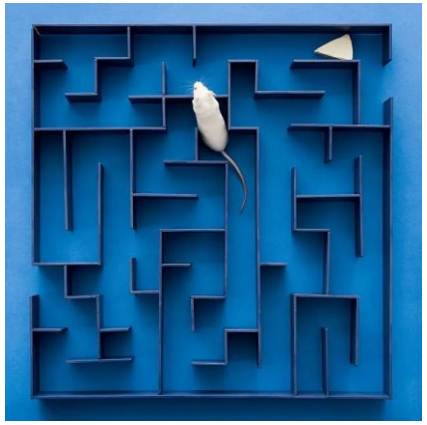
Complexities of our information environments



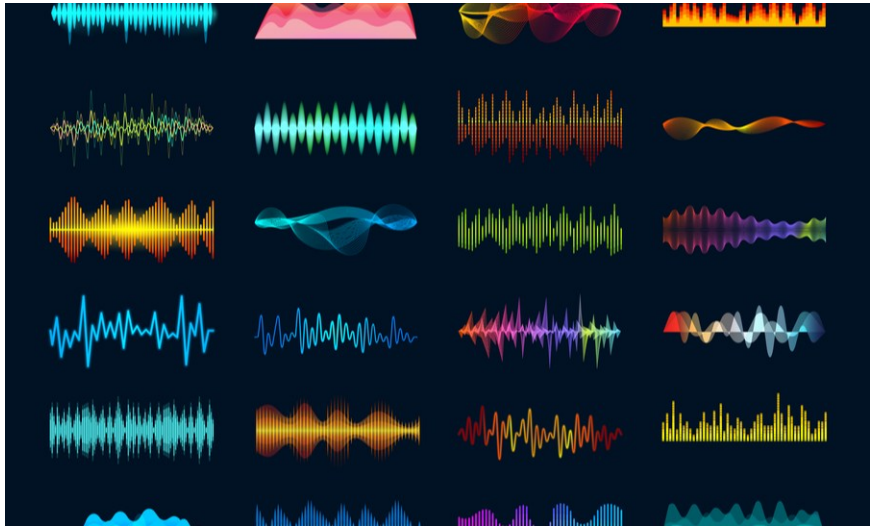
- **Most decision-making situations occur in large worlds: uncertainty and intractability**
- **And most often, time and computational resources are constrained**
- **But we solve remarkably well easy problems of large worlds and good-enough harder problems**
- **The reason is our ability to exploit regularities in information environments**
- **Signal is sovereign: discerning what decreases uncertainties from what cannot**

Information is fundamentally complex



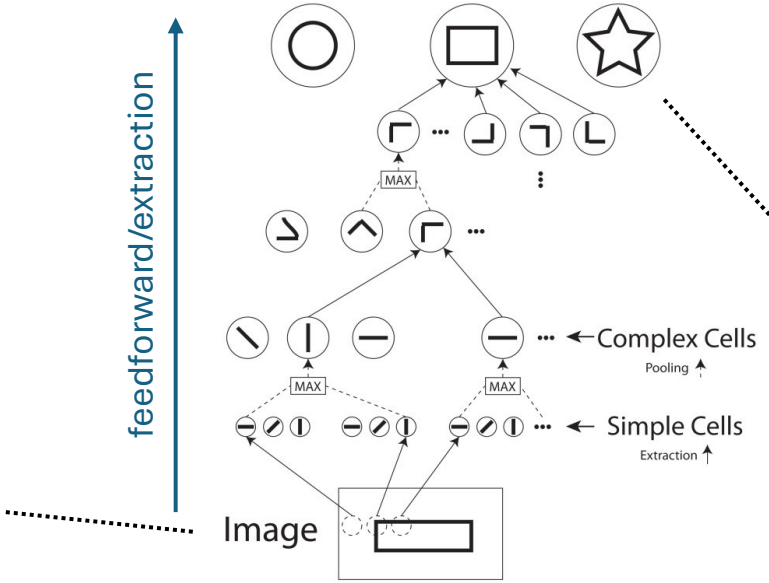


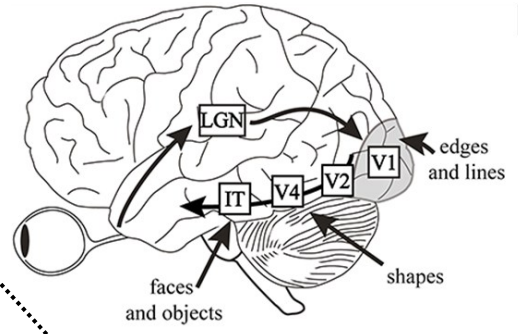
How do process and make sense of low-level information?

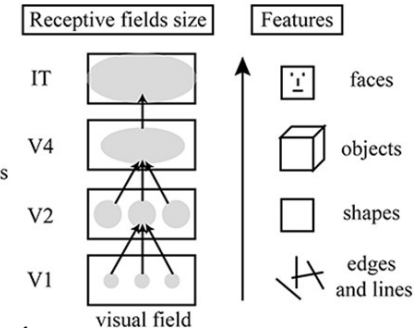


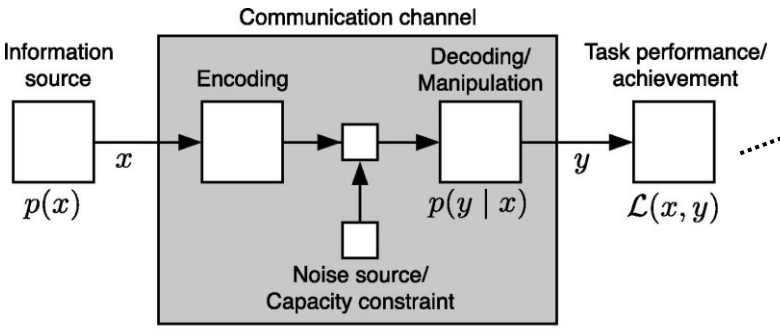
Feature extraction in low-level information

Herzog, Clarke, 2014. *Front. Comput. Neurosci;* de Ladurantaye, Rouat & Vanden-Abeele, 2012. *Visual Cortex…* Zhao et Warren, 2015 ; Fiehler et al., 2019 Sims, 2016. *Cognition*









$p(x)$: Statistics of information source
$p(y \mid x)$: Capacity-limited information channel
$\mathcal{L}(x, y)$: Cost of communication error

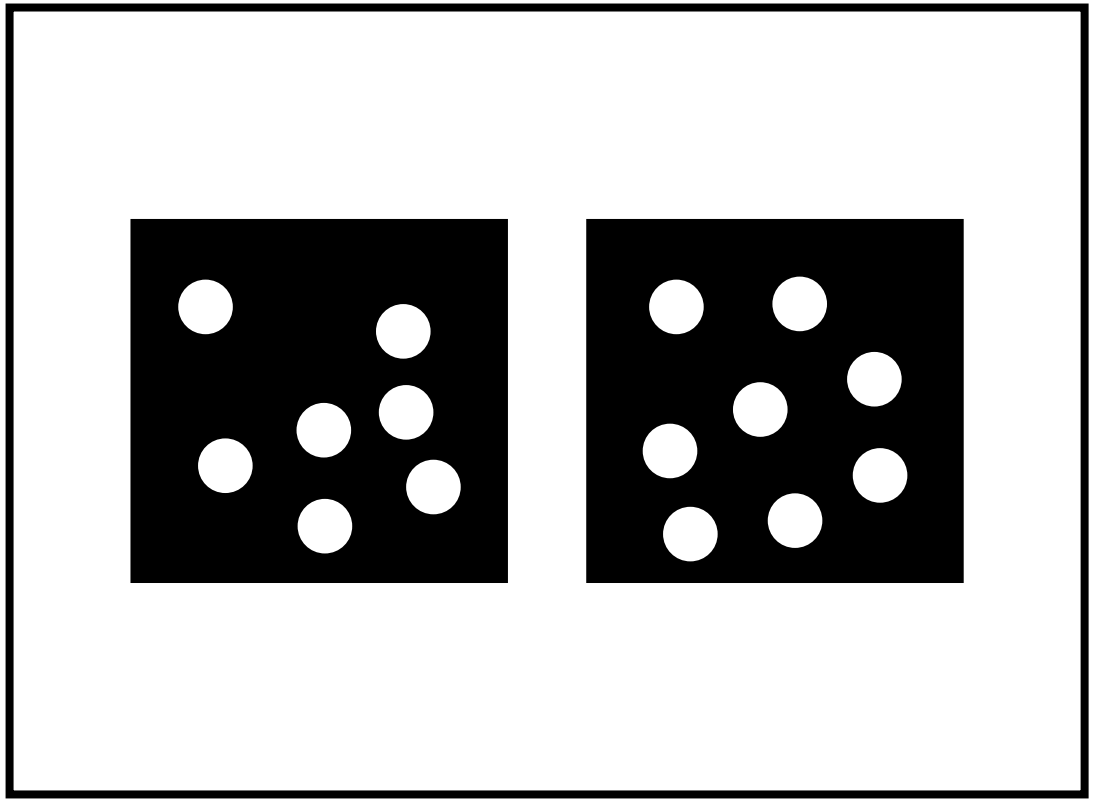
How do we estimate the true state of a low-level information?

High-level: Where is there an open path through the crowd?

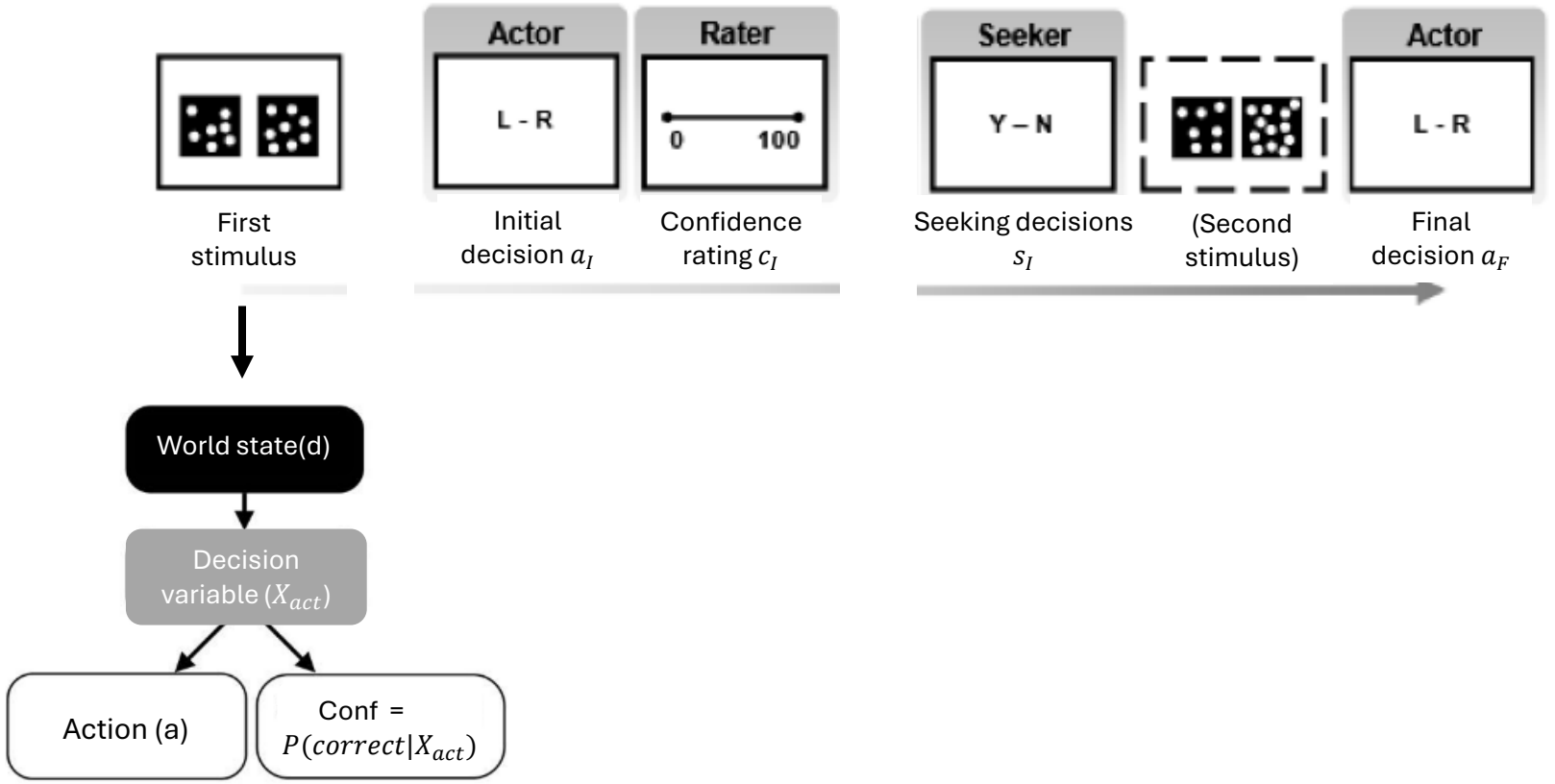
Mid-level: Is there a sufficiently clear corridor on the left side of my position?

Low-level: Where are the edges and boundaries of the walkway? What are pedestrians' directions and speeds?

How do we estimate the true state of a low-level information in the lab?



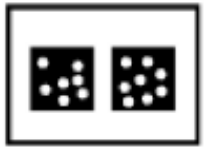
How do we estimate the true state of a low-level information in the lab?



Fleming and Daw, 2017. *Psychological Review*

Schulz et al., 2023. *Psychological Review*

Evidence accumulation and confidence-based choices



First stimulus

Fleming and Daw, 2017 ; Schulz et al., 2023 ;
Fleming, 2024. *Annu. Rev. Of Psy.*

Evidence accumulation and confidence-based choices

Is the corridor on the right side of my position too crowded?

First stimulus: 500ms

Fleming and Daw, 2017 ; Schulz et al., 2023 ;
Fleming, 2024. *Annu. Rev. Of Psy.*

Evidence accumulation and confidence-based choices

Confidence in the orientation

$$p(45° = correct|orientation)$$

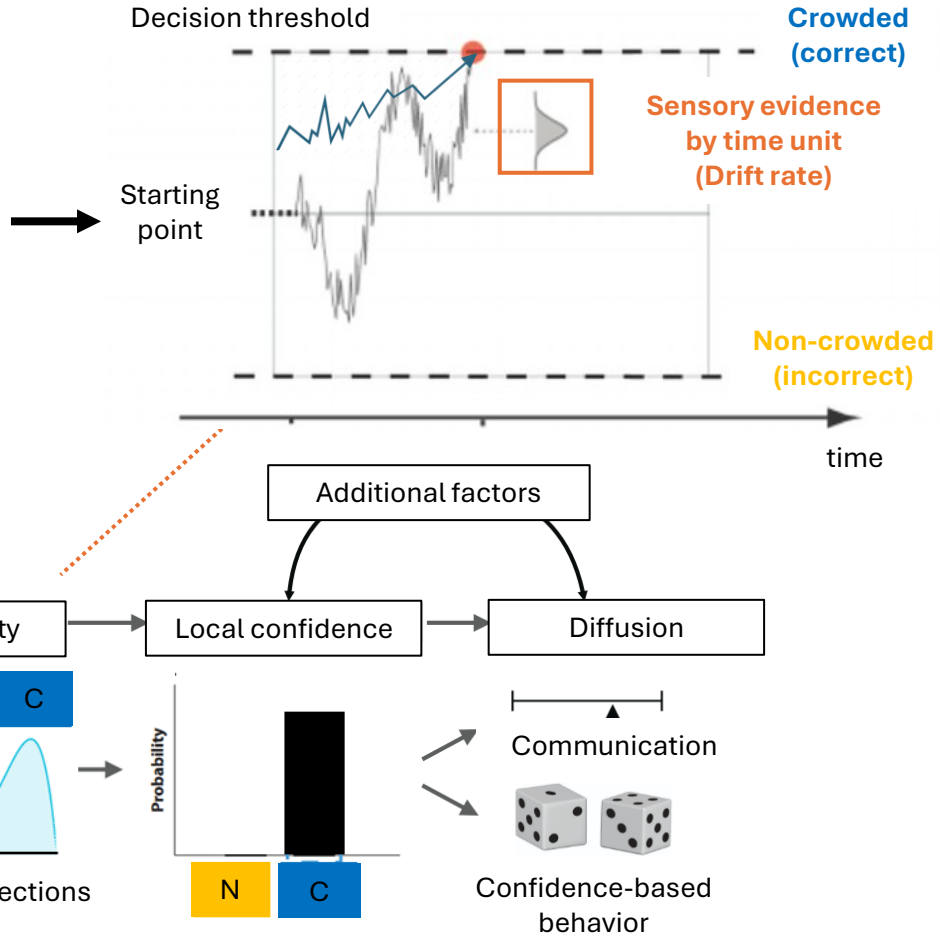
Confidence in the intentions

$$p(keep\ straight = true|street)$$

Confidence in the capacity to traverse

$$p(traverse = aim|trajectory)$$

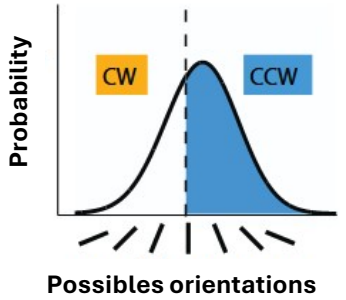
Second stimulus: 250ms



Fleming and Daw, 2017 ; Schulz et al., 2023 ;
Fleming, 2024. *Annu. Rev. Of Psy.*

Evidence accumulation and confidence-based choices

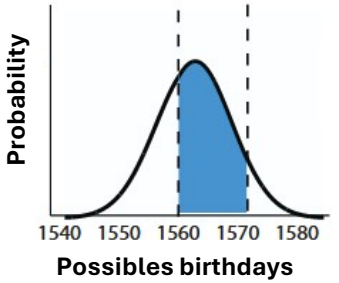
Perception



Propositional confidence in the orientation

$$p(CCW = correct|orientation)$$

Memory / Knowledge



Propositional confidence in the birthday

$$p(1560s = true|birthdays)$$

Competence

Propositional confidence in the capacity to score

$$p(score = aim|trajectory)$$

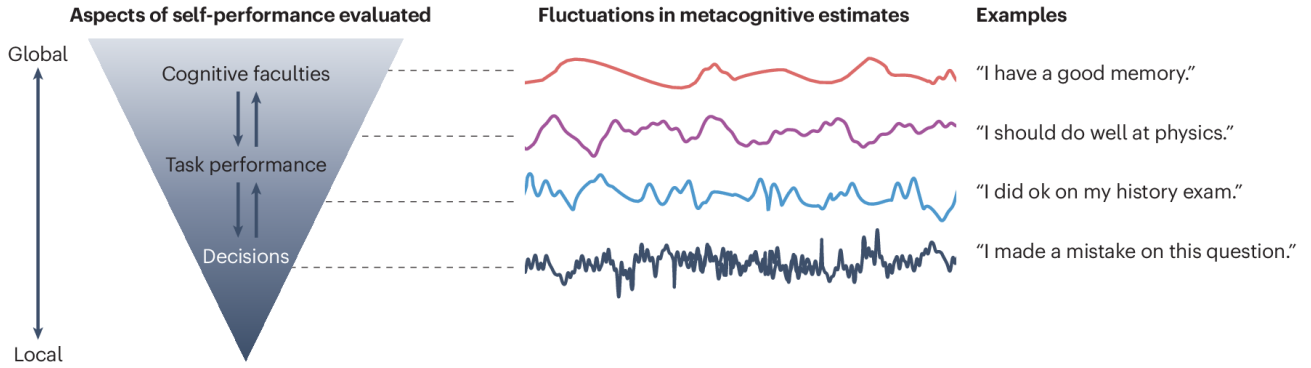
Fleming and Daw, 2017 ; Schulz et al., 2023 ;
Fleming, 2024. *Annu. Rev. Of Psy.*

Metacognitive judgments

Metacognitive judgments estimate **confidence** in one's own decisions, actions, or propositions, across domains and timescales.

It corresponds to the **subjective probability** (belief) that the decision has correctly identified the state of the world.

Metacognition: **set of processes through which an individual forms beliefs about their own mental operations** and evaluates the effectiveness of their cognitive actions.

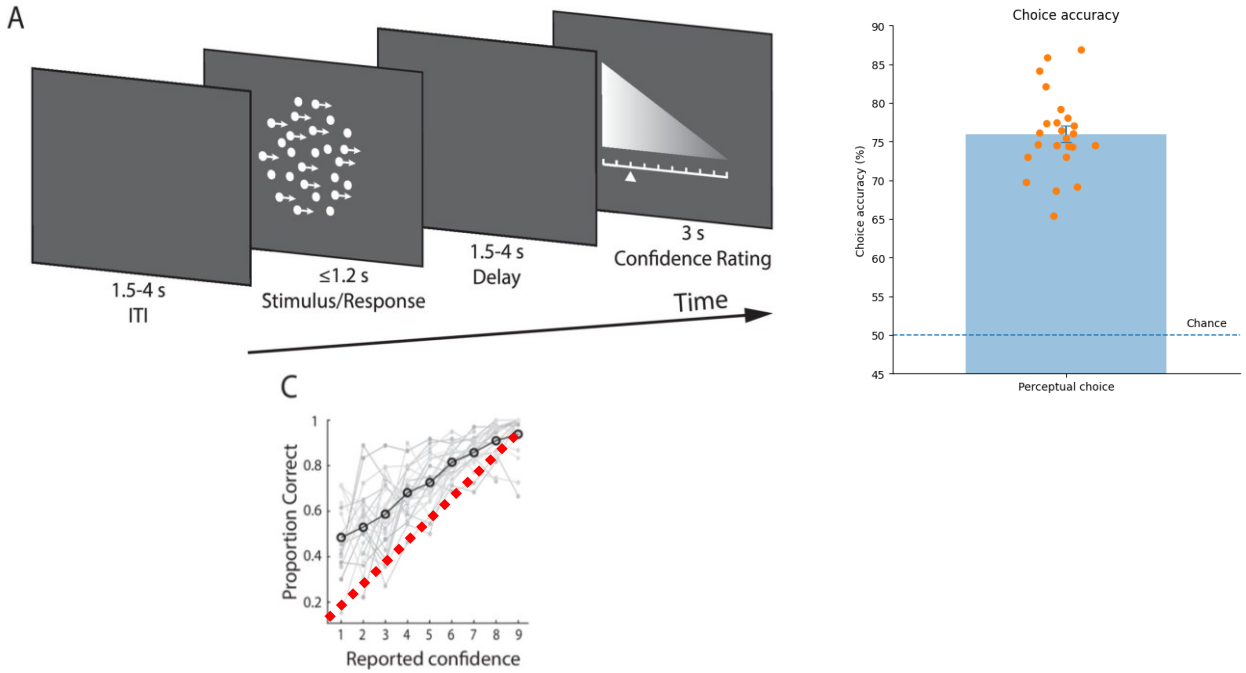


Main functions

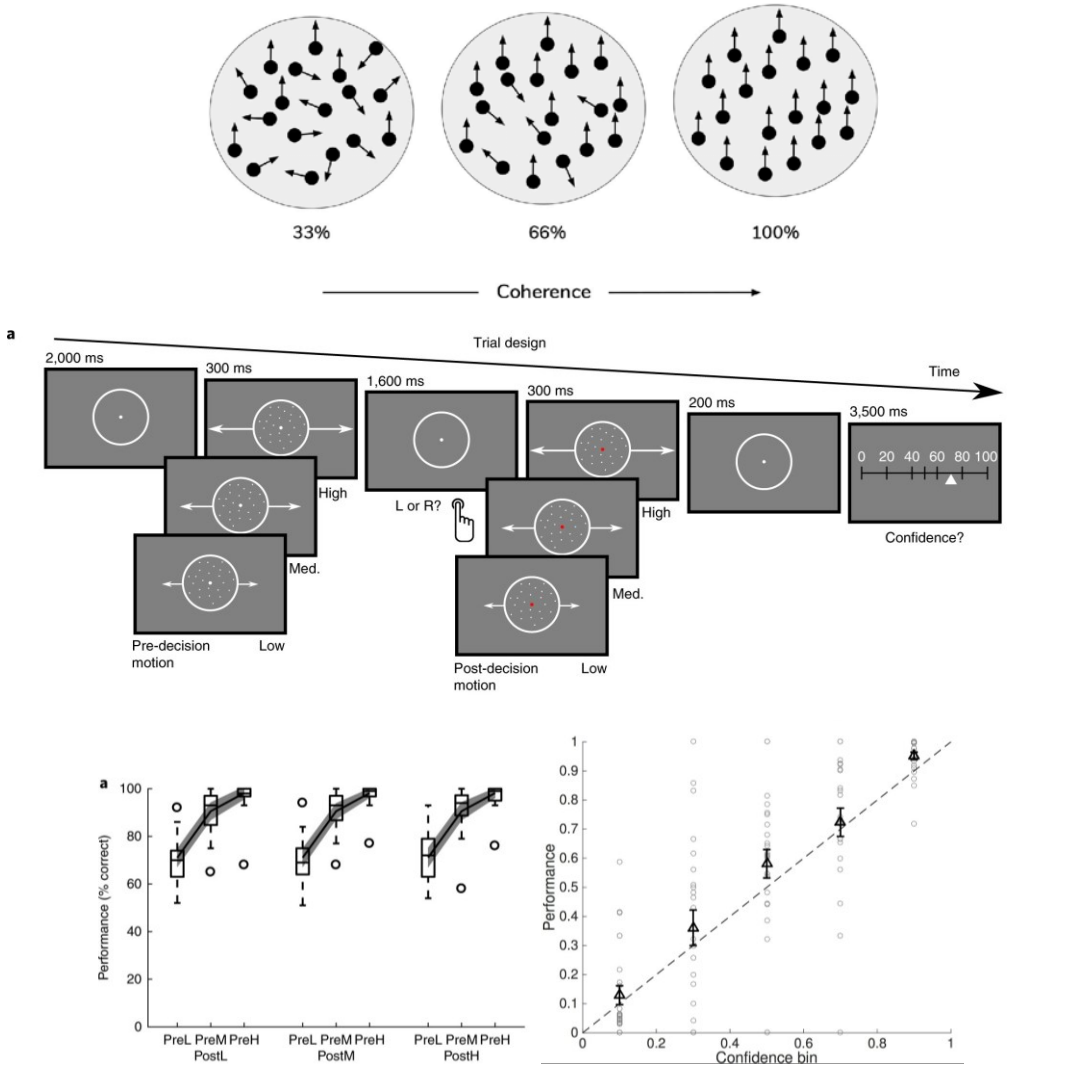
- **Monitoring**: evaluating the state of a mental process (e.g., "Am I confident in my answer?")
- **Control (self-regulation)**: adapting behavior accordingly (e.g., re-evaluation, information seeking, hesitation)
- **Communication**: sharing metacognitive information with others (e.g., expressing confidence, uncertainty, or error awareness to inform or influence others)

Meyniel et al., 2015. *Plos Comp. Bio.*; Pouget et al., 2016. *Nat. Neurosci*; Fleming, 2024; Fleming, 2026. *Nat. Rev. Neurosci.*

How **well** do we process low-level information?



Gherman & Philiastides, 2018. *eLife* (N=24); right figure reconstructed



Fleming et al., 2018. *Nature Neuroscience* (N=25); behavior outside scanner

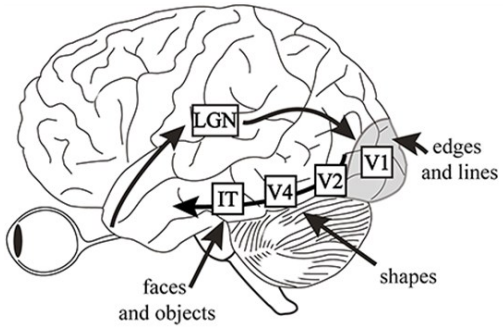
- Humans are great and calibrated at tasks on low-level information like tracking visual vectors
- Performance decreases with information ambiguity

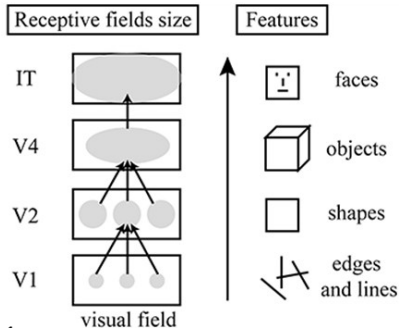
Processing sensory information

Information environment	Proximal input	Core inference	Target	Additional sources of uncertainty
Sensory	Physical signals	What state of the world generated the evidence?	Objects, events, perceptual states	Noise, ambiguity, sensory limitations

How do social and/or mediated information complexify judgments?

Processing social information





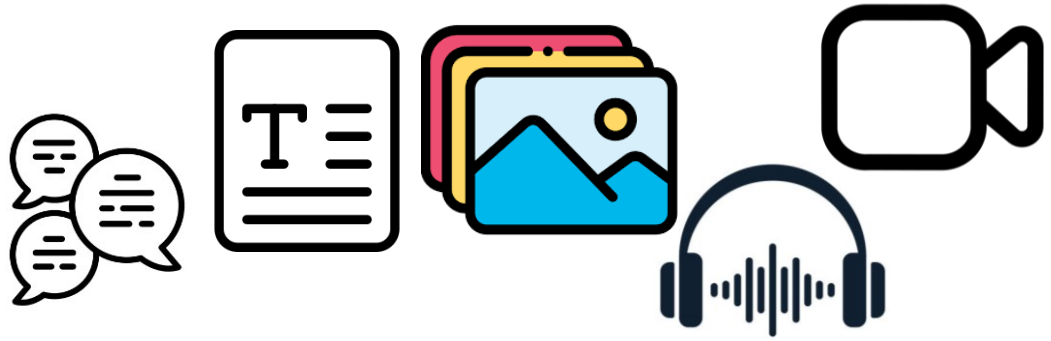
1. Extract low-level perceptual components
2. Extract social components
3. Triggers high-level cognitive processes

Processing social information

Information environment	Proximal input	Core inference	Target	Additional sources of uncertainty
Sensory	Physical signals	What state of the world generated the evidence?	Objects, events, perceptual states	Noise, ambiguity, sensory limitations
Social	Behavior and signals from other agents	What does another person believe, intend…?	Intentions, beliefs, preferences, norms	Strategic behavior, expressive ambiguity, interpretive biases…

Processing mediated information

1. Extract low-level perceptual components
2. Extract social components
3. Triggers high-level cognitive processes



- ***Physical Input -> Output*** **mediated by artefacts** (sound, image, etc.)
- **Encoded through symbolic systems** (language, graphism, etc.)
- Mobilize preliminary resources (prior beliefs) and high-level processes
- Aim at building a common ground by accessing **spatially and temporally indirect realities**

Types of information

Information environment	Proximal input	Core inference	Target	Additional sources of uncertainty
Sensory	Physical signals	What state of the world generated the evidence?	Objects, events, perceptual states	Noise, ambiguity, sensory limitations
Social	Behavior and signals from other agents	What does another person believe, intend…?	Intentions, beliefs, preferences, norms	Strategic behavior, expressive ambiguity, interpretive biases…
Mediated	Symbolic content	What does this representation mean? How much should it be trusted?	Claims, narratives, knowledge, opinions	Source reliability, framing, missing context, manipulation…

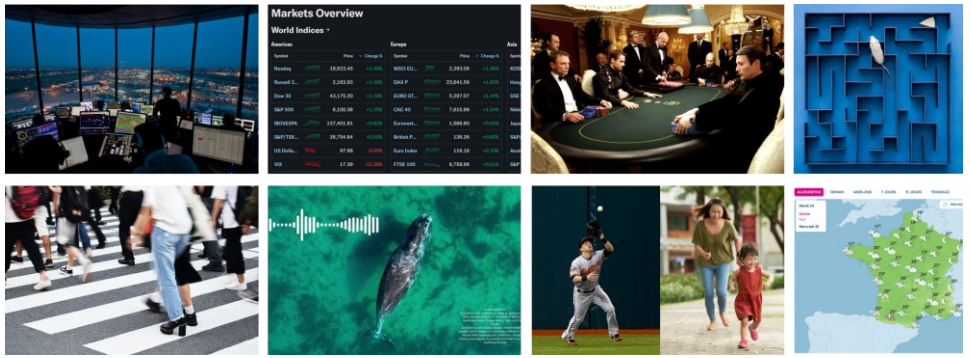
How do we process information in complex information environments?

Propositional judgment on a stimulus

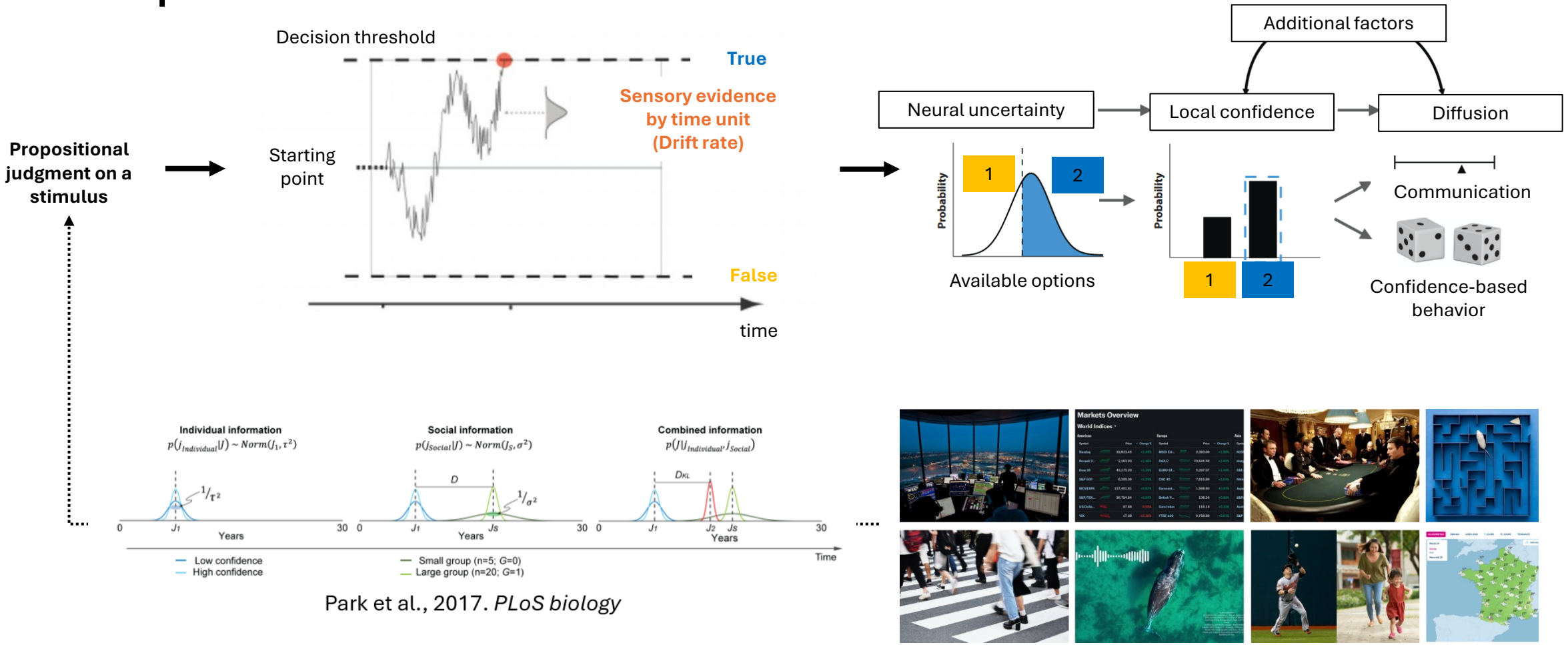
What state of the world generated the evidence?

What does another person believe, intend…?

What does this representation mean?
How much should it be trusted?



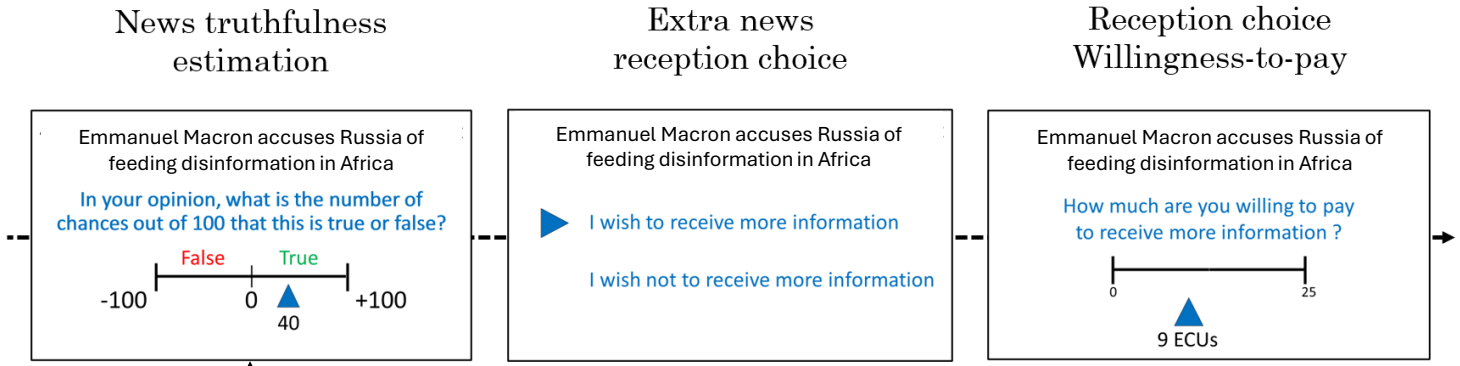
How do we process information in complex information environments?



Park et al., 2017. *PLoS biology*

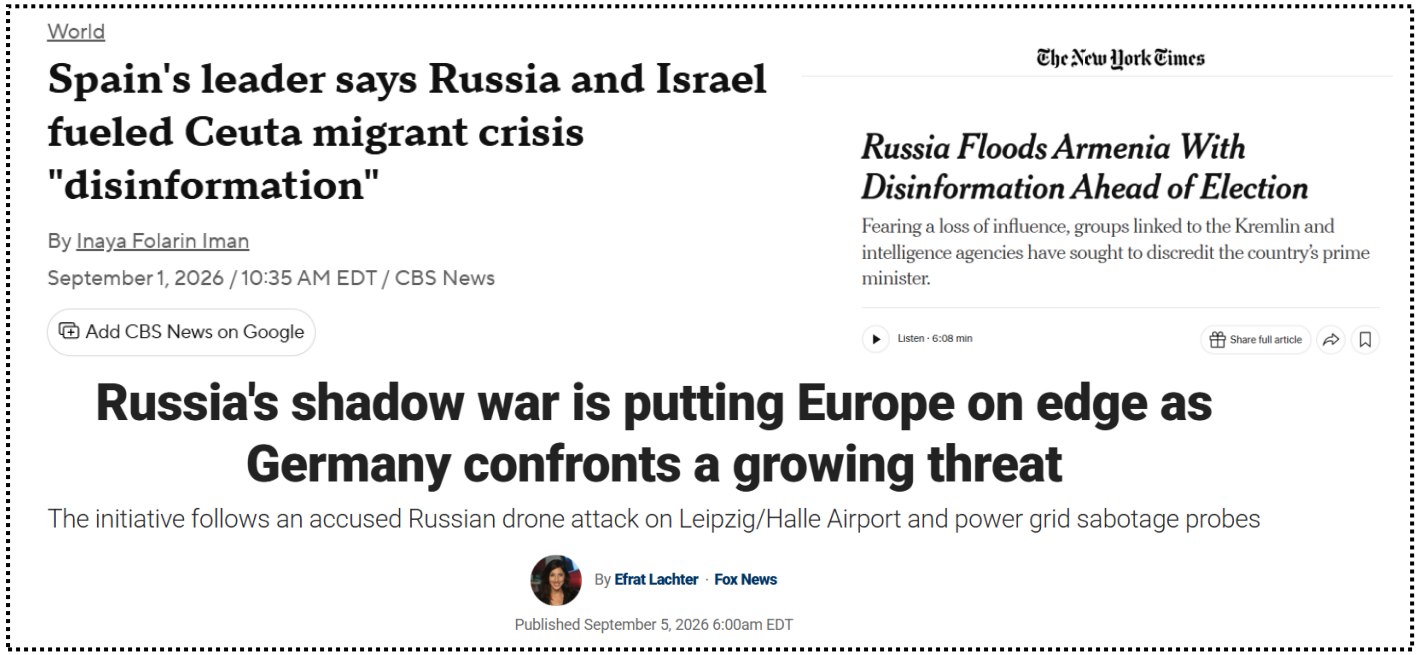
The brain selects then integrates information from multiple sources,
likely using a Bayesian logic,
downweighting the least certain sources (Park et al., 2017; Guigon et al., under review)

How well do we process information in complex information environments?
-> Plausibility estimation of true/false news



Evidence sampling

Beliefs



World

Spain's leader says Russia and Israel fueled Ceuta migrant crisis "disinformation"

By Inaya Folarin Iman

September 1, 2026 / 10:35 AM EDT / CBS News

Add CBS News on Google

The New York Times

Russia Floods Armenia With Disinformation Ahead of Election

Fearing a loss of influence, groups linked to the Kremlin and intelligence agencies have sought to discredit the country's prime minister.

Listen · 6:08 min

Share full article

Russia's shadow war is putting Europe on edge as Germany confronts a growing threat

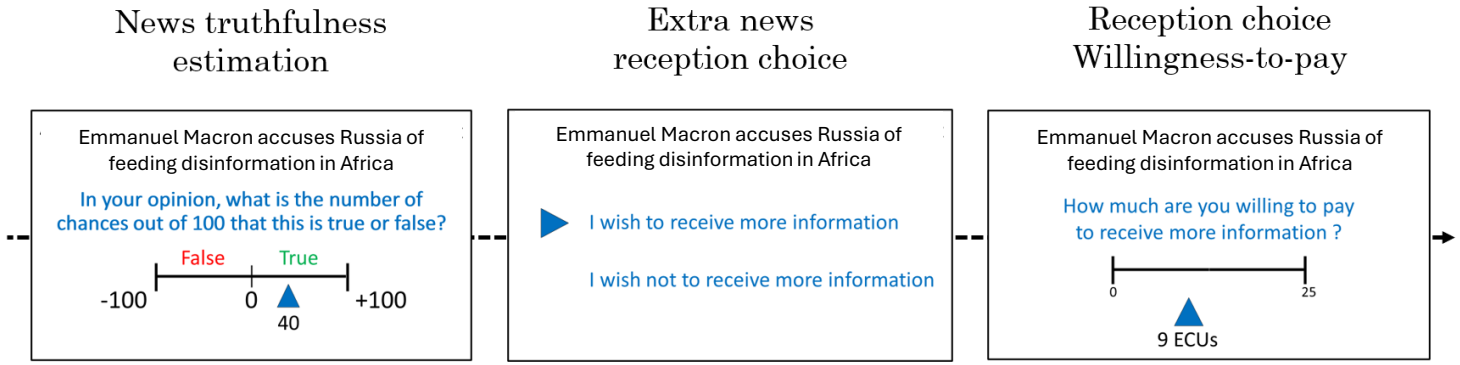
The initiative follows an accused Russian drone attack on Leipzig/Halle Airport and power grid sabotage probes

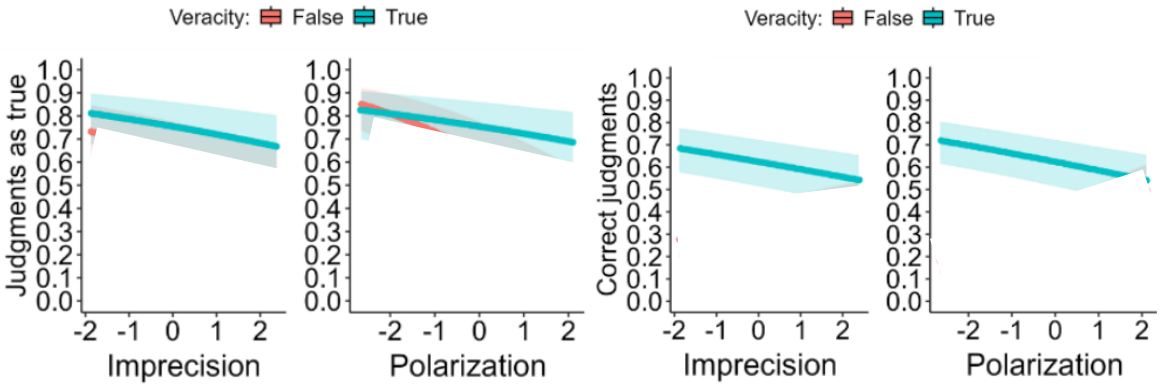
By Efrat Lachter · Fox News

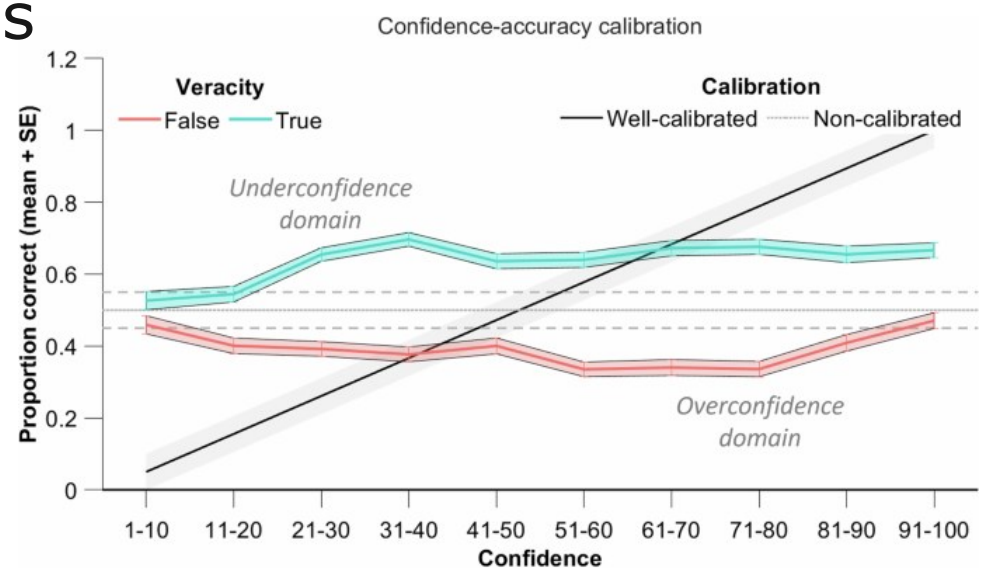
Published September 5, 2026 6:00am EDT

Guigon, Villeval & Dreher, 2024. *Comm. Psychol*

How well do we process information in complex information environments?
-> Plausibility estimation of true/false news

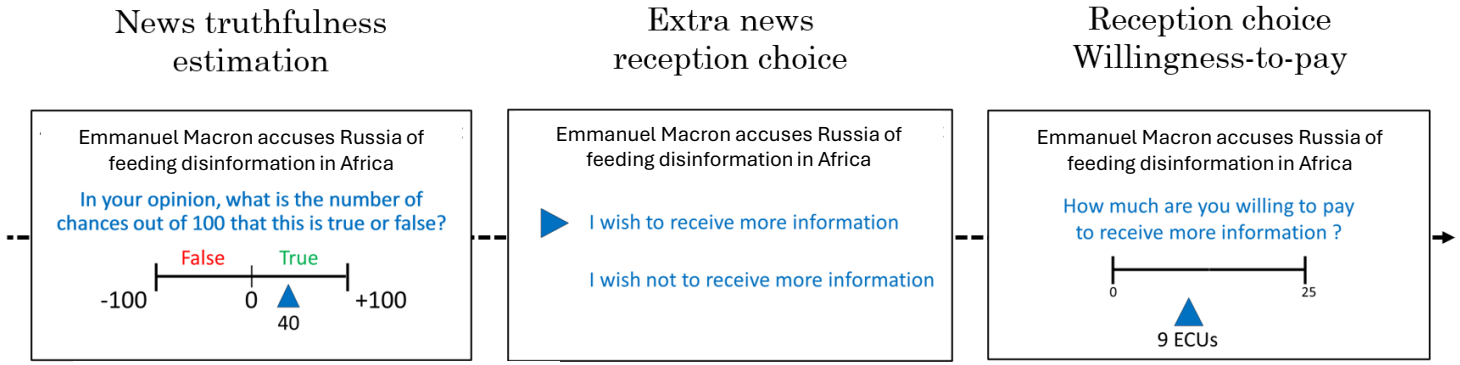


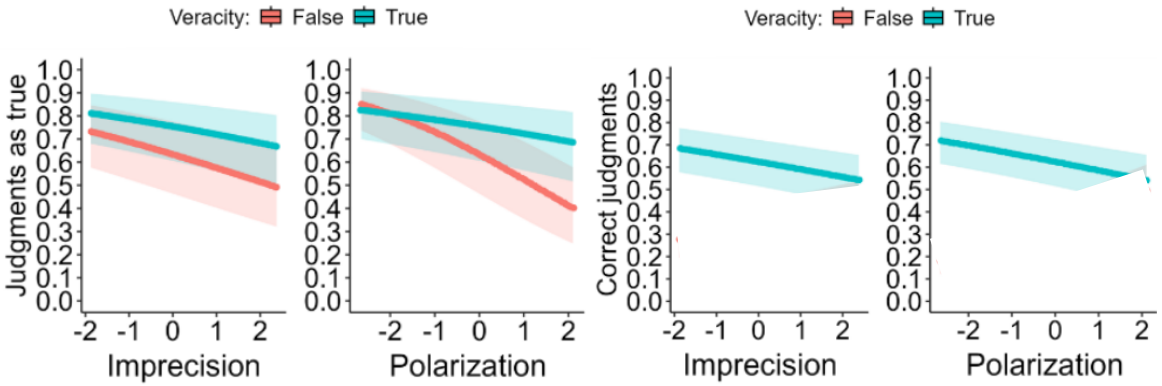


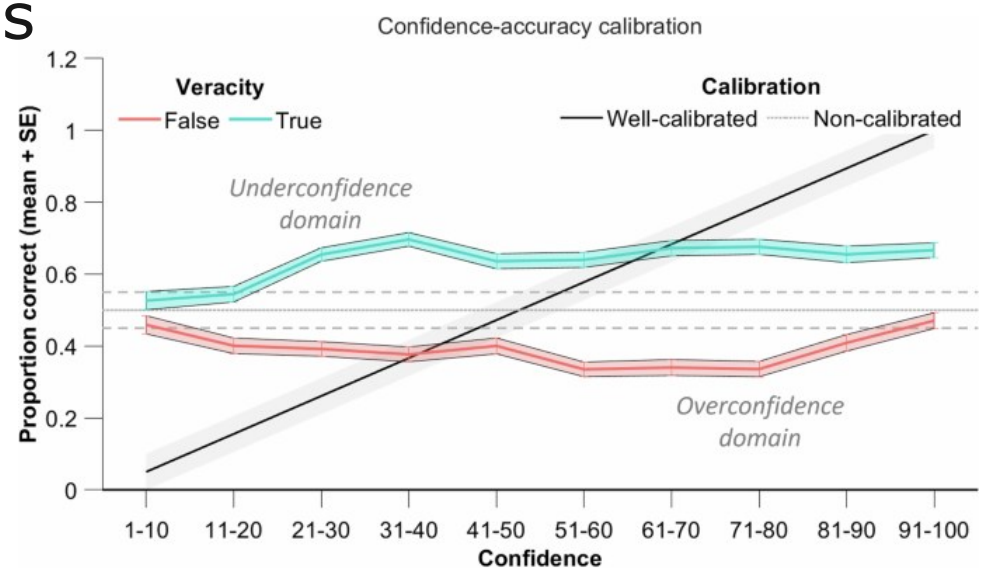


Guigon, Villeval & Dreher, 2024. *Comm. Psychol*

How well do we process information in complex information environments?
-> Plausibility estimation of true/false news

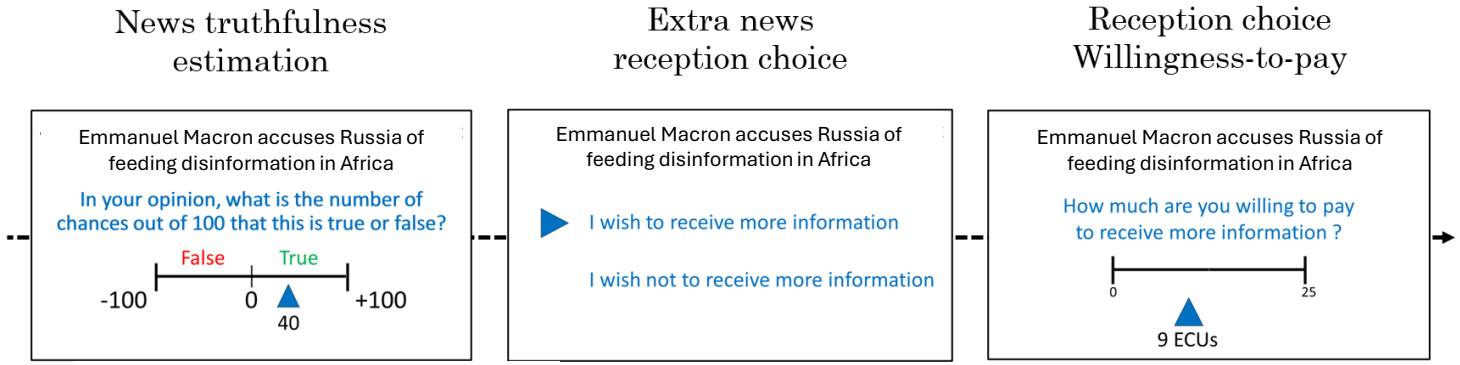


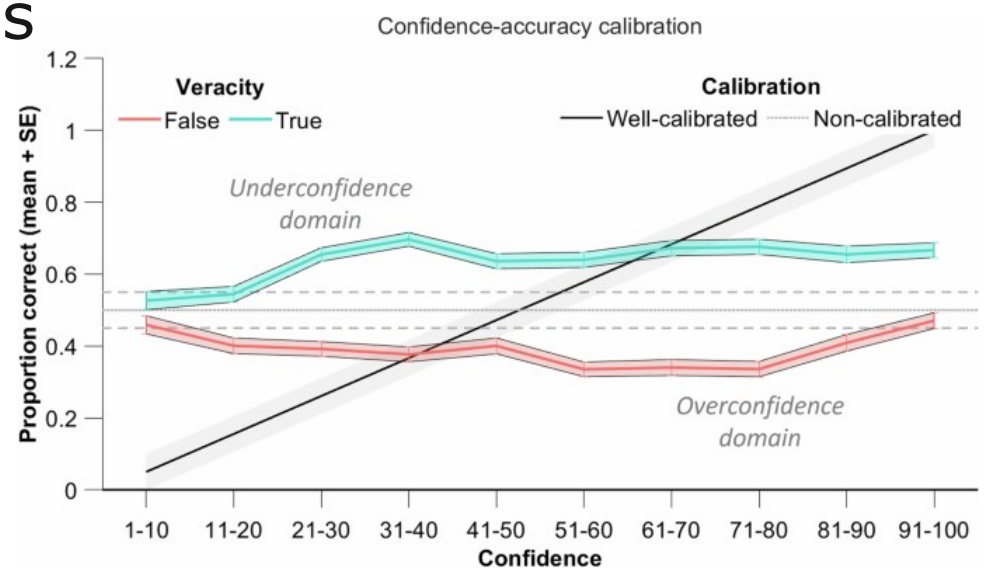


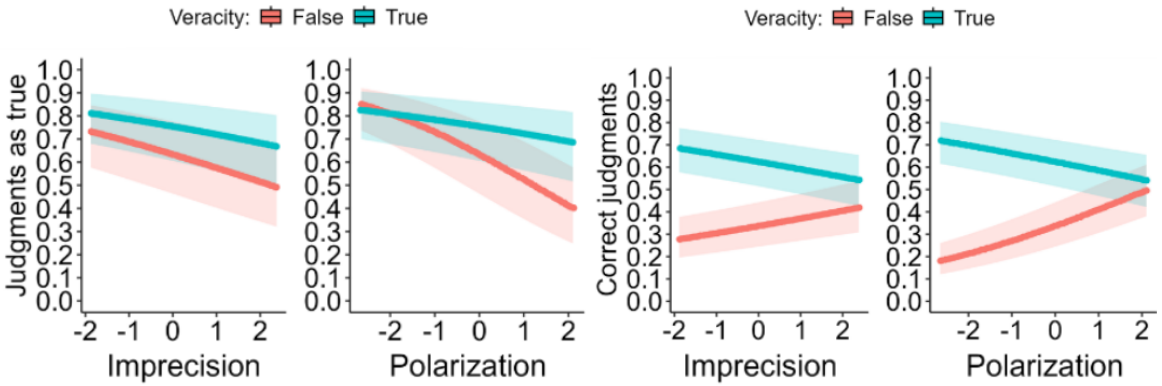


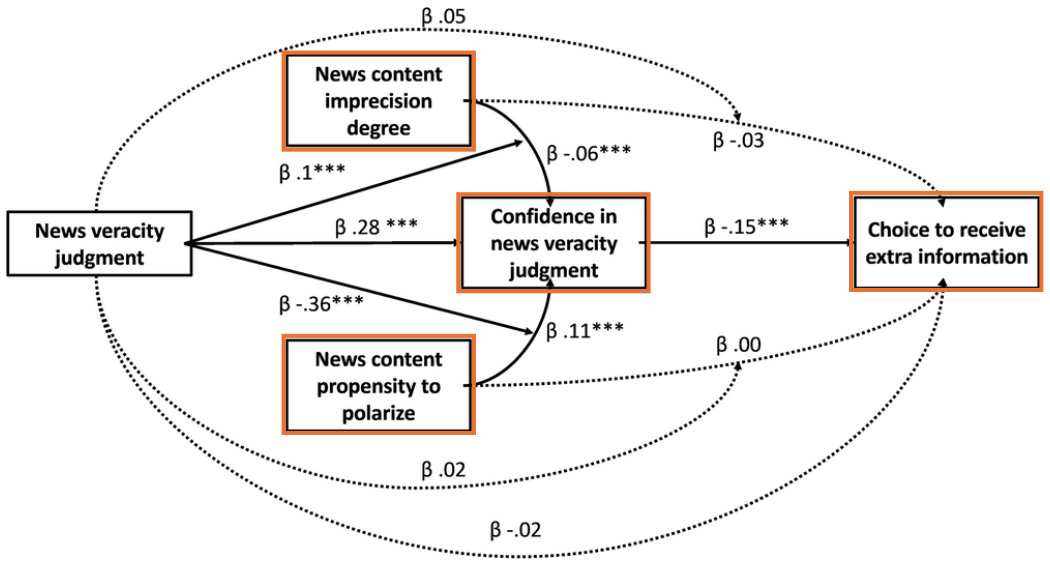
Guigon, Villeval & Dreher, 2024. *Comm. Psychol*

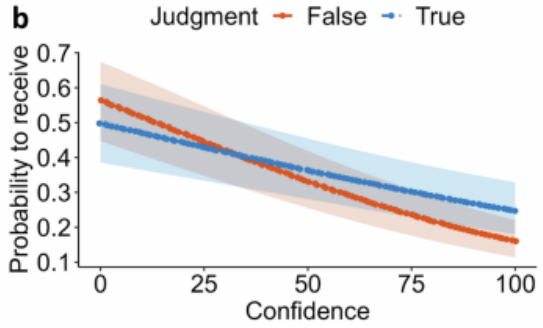
How well do we process information in complex information environments?
-> Plausibility estimation of true/false news











Guigon, Villeval & Dreher, 2024. *Comm. Psychol*

Discernment in tasks with social / mediated information

Individuals are not better than chance at:

- Detecting deception
  (Belot & van de Ven, 2017. *Exp. econ.*;
  Bond & DePaulo, 2006. *Pers. and Soc. Psyc. Review;*
  Bond & DePaulo, 2008. *Psyc. Bulletin*)
- Distinguishing between true and false videos about news events
  (Serra-Garcia & Gneezy, 2021. *Am. Econ. Rev.*)
- Distinguishing between true and false news
  (Guigon, Villeval & Dreher, 2024. *Commun Psychol)*

Individuals are worse than chance at detecting false income reports
(Dwenger & Lohse, 2019. *Joep*; Konrad et al., 2014. *Jebo*)

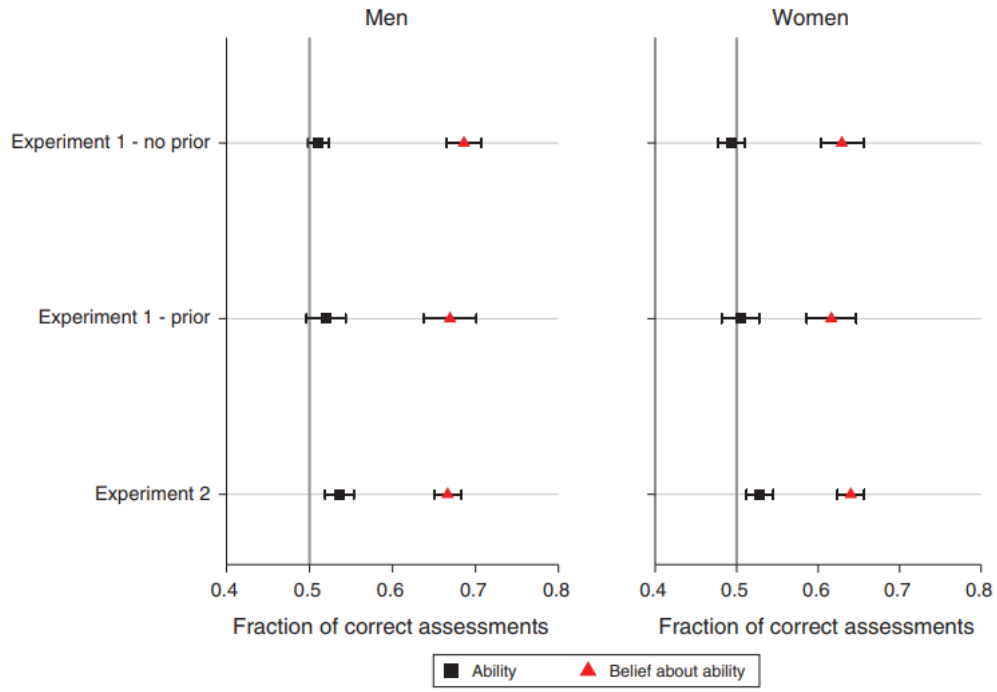


FIGURE 1. ABILITY AND ABSOLUTE OVERCONFIDENCE, BY EXPERIMENT AND GENDER

Serra-Garcia & Gneezy, 2021

Discernment decreases with task difficulty

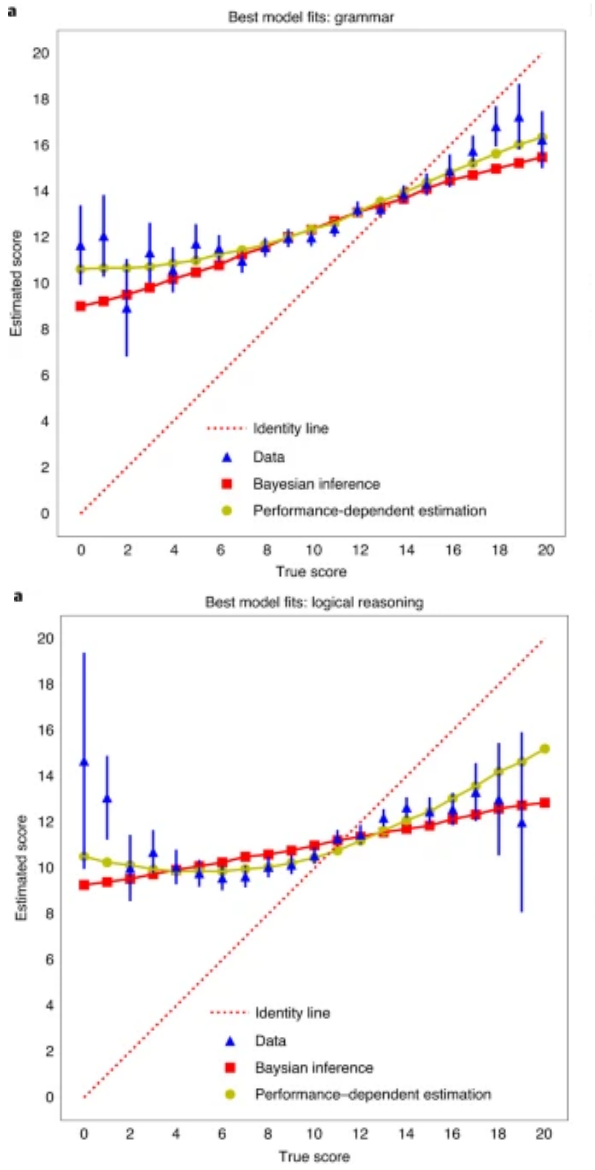
Individuals are not better than chance at:

- Detecting deception
  (Belot & van de Ven, 2017. *Exp. econ.*;
  Bond & DePaulo, 2006. *Pers. and Soc. Psyc. Review;*
  Bond & DePaulo, 2008. *Psyc. Bulletin*)
- Distinguishing between true and false videos about news events
  (Serra-Garcia & Gneezy, 2021. *Am. Econ. Rev.*)
- Distinguishing between true and false news
  (Guigon, Villeval & Dreher, 2024. *Commun Psychol)*

Individuals are worse than chance at detecting false income reports
(Dwenger & Lohse, 2019. *Joep*; Konrad et al., 2014. *Jebo*)

As task difficulty increases, confidence is decreasingly predicting judgment accuracy
(Weber & Brewer, 2004. *J. Exp. Psychol. Appl.;*
Moore & Healy, 2008. *Psychol. Rev.;* Moore & Schatz, 2017. *Soc. Pers. Psychol.;*
Boldt et al., 2017. *J. Exp. Psychol. Hum. Percept. Perform.)*

Correspondance between judgments and real performance



Jansen et al., 2021, ill calibration (population-level)

Summary

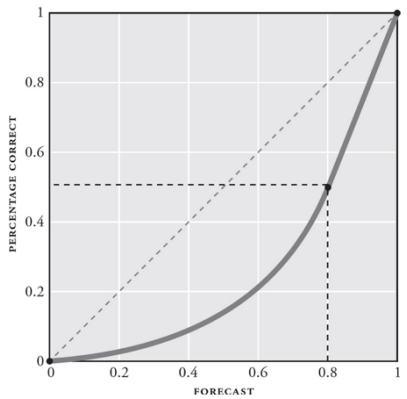
Information environment	Proximal input	Core inference	Target	Additional sources of uncertainty
Sensory	Physical signals	What state of the world generated the evidence?	Objects, events, perceptual states	Noise, ambiguity, sensory limitations
Social	Behavior and signals from other agents	What does another person believe, intend...?	Intentions, beliefs, preferences, norms	Strategic behavior, expressive ambiguity, interpretive biases...
Mediated	Symbolic content	What does this representation mean? How much should it be trusted?	Claims, narratives, knowledge, opinions	Source reliability, framing, missing context, manipulation...

- Most our judgments/decisions occur under complexity: uncertainty, intractability, abundant or noisy information, with time and computational constraints

- The more complex, the more levels of judgment we are required to perform
- Leading eventually to adopt *good-enough* solutions (cf. *heuristics & biases*)

- Metacognition guides re-evaluation, information-seeking, etc.
- Complexity decreases the relationship between accuracy and metacognition *(how much one knows how much they know)*

Curb complexity-related issues:



- Improving metacognition
- Improving information environments

Rethinking misinformation through plausibility estimation and confidence calibration

Valentin Guigon, Lucille Geay & Caroline J. Charpentier

Communications Psychology **4**, Article number: 24 (2026) | Cite this article

Metacognitive calibration is a critical and understudied lever for strengthening resilience in complex information ecosystems. The central challenge is not only that misinformation spreads, but that human judgments are formed under bounded resources and multiple goals, making the estimation of a claim's plausibility a primary cognitive bottleneck. Addressing misinformation, therefore, requires interventions that act on this evaluative process, rather than relying solely on binary truth detection. Metacognitive training targets this neglected dimension and thus represents one promising route for complementing content-level approaches in the fight against misinformation.

Crucial to study how to improve people's ability to make judgments, with investigating metacognitive training

LLMs as Oracles: Reliance on LLMs for Subjective Personal Questions

Myra Cheng*
myra@cs.stanford.edu
Stanford University
Stanford, California, USA

Lujain Ibrahim*
lujain.ibrahim@oii.ox.ac.uk
University of Oxford
Oxford, England

Grace Liu
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA

Michelle S. Lam
Stanford University
Stanford, California, USA

Vishakh Padmakumar
Stanford University
Stanford, California, USA

Nick Madibekov
Stanford University
Stanford, California, USA

Diyi Yang
Stanford University
Stanford, California, USA

Dan Jurafsky
Stanford University
Stanford, California, USA

Figure 1: While people have long consulted oracles for subjective personal questions, these oracles were limited in access and specificity. Now, people turn to LLMs for similar questions without these constraints, raising the risk of excessive reliance. We build a typology of the ways people use LLMs as oracles, from offloading beliefs and normative judgments to offloading decisions and actions. Using this typology, we find that LLM-as-oracle use is increasing across multiple real-world usage datasets, and that people are often unaware of their own LLM-as-oracle use. We also show how model behavior and people's perceptions of AI drive this reliance. Finally, we propose interventions on both sides: how people can use LLMs in alternative, non-oracle ways, and how model outputs can better facilitate users' autonomy.

Abstract

We characterize how people are turning to LLMs as *oracles*: all-knowing authorities on subjective personal questions. Motivated by risks to users' autonomy and well-being, we develop a typology and LLM-based methods to measure this form of AI reliance at scale and understand how people are offloading judgment and decision-making to AI. Applying our typology to public usage data (68K prompts from WildChat and ThoughtTrace), we find that LLM-as-oracle use has increased over time (2023–2026) and is more prevalent among younger users. We further build a privacy-preserving data donation tool to analyze individuals' longitudinal usage data (140K prompts from 52 participants), identifying similar trends. People are often unaware of their own LLM-as-oracle use, and express dissatisfaction with this behavior after seeing our tool's analysis. Finally, we identify two drivers of LLM-as-oracle use: people's perceptions of AI and the behavior of AI models themselves, which motivate possible interventions to support users' self-deliberation.

*Both authors contributed equally to this research.

CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI**; **Natural language interfaces**; • **Computing methodologies** → **Natural language generation**.

Keywords

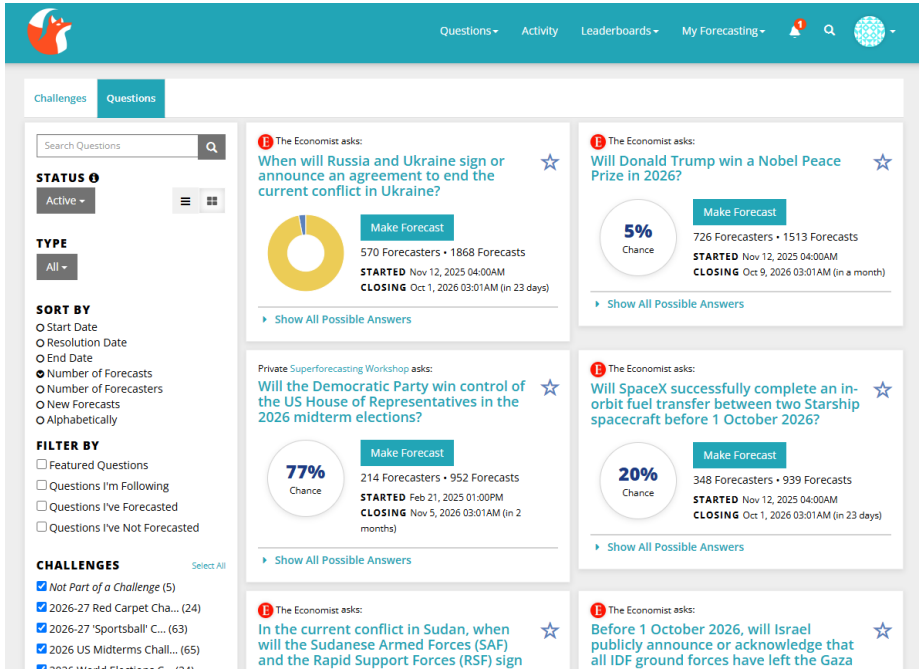
artificial intelligence, large language models, AI reliance, personal advice, autonomy, offloading, perceptions of AI

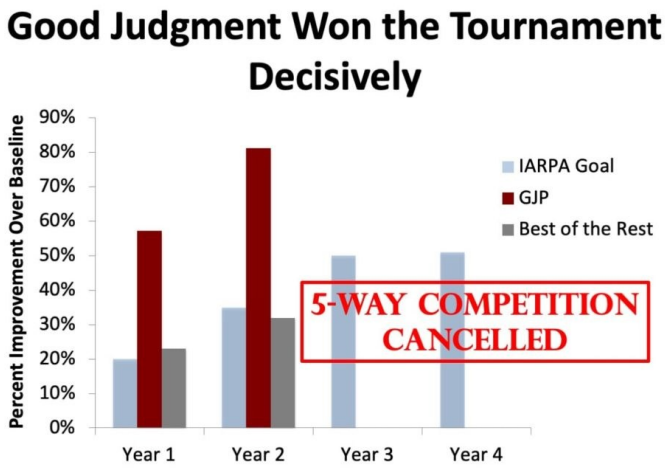
How are human judgments expected to transform with AI?

I. Can AI make accurate estimations of the state of the world?

(Super) forecasting

See Mellers et al., 2014. *Psychological Science,*
Mellers et al., 2015. *Journal of Experimental Psychology: Applied*





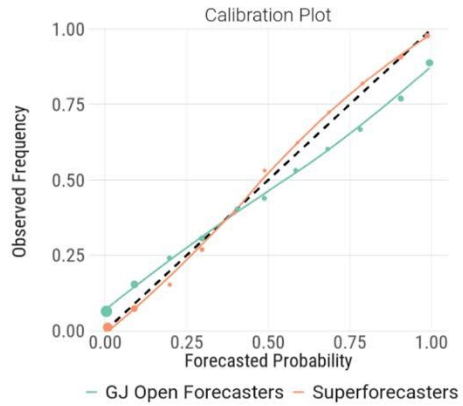


Figure 7. Calibration curves of each forecasting group.
Source: Good Judgment Inc

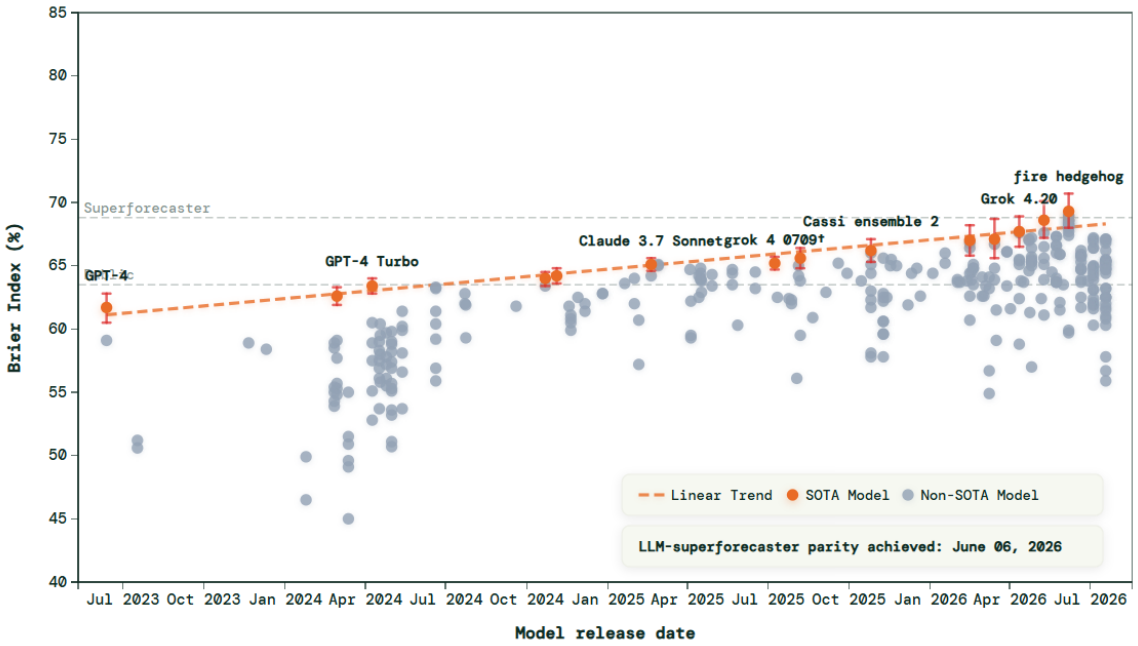
Combine intellectual processes based on niche knowledge, probabilistic judgment, and performant crowds to estimate probabilities of **highly uncertain events**

1. **Triage** (avoid wasting time on irrelevant problems)
2. **Break problems down**
3. **Balance inside and outside views** (identify comparison classes)
4. **Update beliefs** (Bayesian updating + calibrate confidence)
5. **Remain open to being wrong**
6. **Reduce uncertainty** (distinguish 60/40 from 55/45)
7. **Balance caution and decisiveness**
8. **Learn from failures and successes**
9. **Team management** (stepping back, precise questioning, etc.)
10. **Balance opposing errors**

Philip Tetlock & Dan Gardner, 2016, in *Superforecasting*

AI forecasting

Projected LLM-superforecaster parity on *ForecastBench*.



Base model LLM *without additional tools*

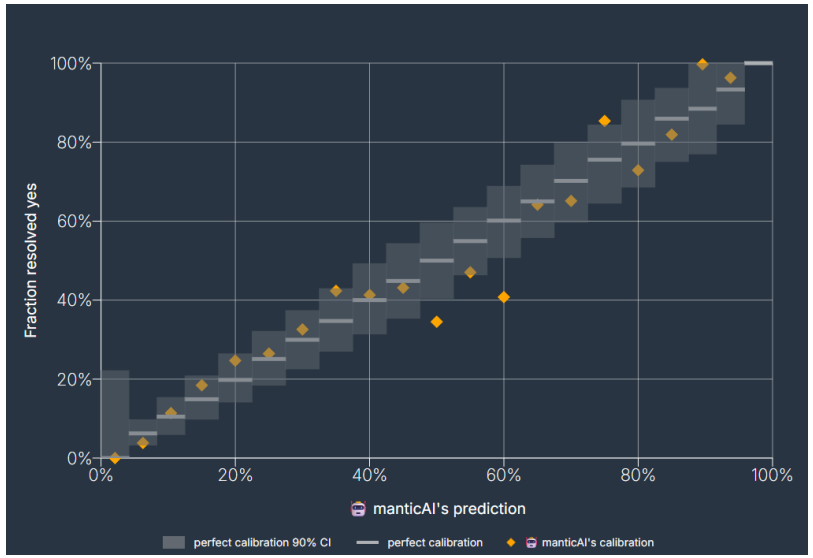
Rank	Org	Model	Overall
1		Superforecaster median forecast	67.8
2		Public median forecast	62.7
3		claude-sonnet-4-6-adaptive-thinking-16000	62.6
4		o3-2025-04-16-scratchpad	61.8
5		claude-sonnet-5-adaptive-thinking-16000	61.6
6		minimax-m3-adaptive-thinking-12000	61.2
6		grok-4.20-0309-reasoning	61.2
8		gpt-5.5-2026-04-23	61.1
9		claude-opus-4-1-20250805	61.0
10		claude-3-7-sonnet-20250219-scratchpad	60.9

Showing 1 to 10 of 129 entries
last updated 2026-09-10
Previous Next

LLM *with additional tools*

Rank	Org	Model	Overall
1		fire hedgehog	69.3
2		ceramic-kettle	69.0
3		blue croc	68.9
4		Superforecaster median forecast	68.8
5		Cassi-2026-05-10	68.6
5		big green leaf	68.6
5		captain-jack	68.6
5		rice-demon	68.6
9		dragon-brother	68.5
10		iron-compass	68.3

Showing 1 to 10 of 302 entries
last updated 2026-09-10
Previous Next



manticAI calibration plot on *Metaculus* (as of February 2026)

AI systems can make accurate estimations of the state of a large world

How are human judgments expected to transform with AI?

II. How is AI changing the process by which humans form judgments?

We are incentivized to pair with AI for making complex judgments

AI-Augmented Predictions: LLM Assistants Improve Human Forecasting Accuracy

PHILIPP SCHOENEGGER, LSE, London, United Kingdom of Great Britain and Northern Ireland
PETER S. PARK, MIT, Cambridge, MA, USA
EZRA KARGER, Federal Reserve Bank of Chicago, Chicago, IL, USA
SEAN TROTT, University of California San Diego, San Diego, CA, USA
PHILIP E. TETLOCK, University of Pennsylvania, Philadelphia, PA, USA

SCIENCE ADVANCES | RESEARCH ARTICLE

COMPUTER SCIENCE

Wisdom of the silicon crowd: LLM ensemble prediction capabilities rival human crowd accuracy

Philipp Schoenegger[1]*, Indre Tuminauskaite[2], Peter S. Park[3], Rafael Valdece Sousa Bastos[4], Philip E. Tetlock[5,6]

Front. Artif. Intell., 12 January 2026
Sec. AI in Finance
Volume 8 - 2025 | https://doi.org/10.3389/frai.2025.1696423

Artificial intelligence in financial market prediction: advancements in machine learning for stock price forecasting

Cooperation → Delegation

AI can harm knowledge, preferences, and the sense of confidence

PNAS RESEARCH ARTICLE | PSYCHOLOGICAL AND COGNITIVE SCIENCES

Human preferences are susceptible to covertly misaligned AI advice

Sahand Sabour[a,1,2], June M. Liu[b,c,1,2], Siyang Liu[d], Chris Z. Yao[a], Shiyao Cui[a], Wen Zhang[e], Xuanming Zhang[a], Yaru Cao[a,f], Advait Bhat[g], Jian Guan[h], Wei Wu[h], Rada Mihalcea[d], Hongning Wang[a], Tim Althoff[e], Tatia M. C. Lee[b,c,2], and Minlie Huang[a,2]

Affiliations are included on p. 8.

Edited by Susan T. Fiske, Princeton University, Princeton, NJ; received January 29, 2026; accepted August 18, 2026

AI assistants are increasingly used as advisors to guide decisions, yet little is known about how people evaluate such advice when the advisor's underlying intent conflicts with their interests. We examine how covert misalignment shapes choices in a randomized experiment ($N = 233$ participants; 699 observations) in which participants rated financial or emotional decisions before and after consulting one of three AI advisors: a neutral advisor, a misaligned advisor with a hidden objective to promote an inferior option, or a strategy-enhanced misaligned advisor additionally equipped with established tactics of covert influence. Across both domains, exposure to misaligned advisors shifted preferences away from optimal options and toward inferior alternatives, increasing the odds of preferring the incentivized (inferior) option over the optimal option by ≈5 to 8 times (up to +38 percentage points). Adding explicit influence strategies did not reliably strengthen these effects. Notably, participants continued to rate misaligned advisors as helpful, revealing a systematic disconnect between susceptibility to misaligned advice and subjective evaluations of advisor quality. These findings have implications for the design and governance of AI-mediated advice.

Significance

Many people are turning to AI for advice that shapes their decisions. Yet, they cannot see the incentives behind such advice and typically assume it aligns with their interests. But what happens when an advisor has hidden objectives that push toward inferior alternatives? In a randomized experiment spanning financial and emotional contexts, we found that covertly misaligned advisors shifted

PNAS RESEARCH ARTICLE | ECONOMIC SCIENCES

Generative AI without guardrails can harm learning: Evidence from high school mathematics

Hamsa Bastani[a,b,1], Osbert Bastani[c,1], Alp Sungu[a,1,2], Haosen Ge[b], Özge Kabakcı[d], and Rei Mariman[e]

Affiliations are included on p. 7.

Edited by Emma Brunskill, Stanford University, Stanford, CA; received November 3, 2024; accepted May 5, 2025 by Editorial Board Member Mark Granovetter

Generative AI is poised to revolutionize how humans work, and has already demonstrated promise in significantly improving human productivity. A key question is how generative AI affects learning—namely, how humans acquire new skills as they perform tasks. Learning is critical to long-term productivity, especially since generative AI is fallible and users must check its outputs. We study this question via a field experiment where we provide nearly a thousand high school math students with access to generative AI tutors. To understand the differential impact of tool design on learning, we deploy two generative AI tutors: one that mimics a standard ChatGPT interface ("GPT Base") and one with prompts designed to safeguard learning ("GPT Tutor"). Consistent with prior work, our results show that having GPT-4 access while solving problems significantly improves performance (48% improvement in grades for GPT Base and 127% for GPT Tutor). However, we additionally find that when access is subsequently taken away, students actually perform worse than those who never had access (17% reduction in grades for GPT Base)—i.e., unfettered access to GPT-4 can harm educational outcomes. These negative learning effects are largely mitigated by the safeguards in GPT Tutor. Without guardrails, students attempt to use GPT-4 as a "crutch" during practice problem sessions, and subsequently perform worse on their own. Thus, decision-makers must be cautious about design choices underlying generative AI deployments to preserve skill learning and long-term productivity.

Significance

While generative AI has been shown to enhance productivity, its influence on learning new skills remains unclear. Our research examines the impact of generative AI, specifically GPT-4, on student learning in math education. Through a large-scale field experiment in a high school, our study reveals that although AI-based tutoring improves performance during practice sessions, students relying on the technology may underperform when access to AI is subsequently

AI advice suppresses people's willingness to say "I don't know", even when the advice is wrong and accuracy is incentivized

Chiara Marcoccia[1], Walter Quattrociocchi[2], Valerio Capraro[3]

[1] Ecole Normale Supérieure, Paris, France

[2] University of Rome "La Sapienza", Rome, Italy

[3] University of Milan-Bicocca, Milan, Italy

Contact author: valerio.capraro@unimib.it

Abstract

Knowing when to say "I don't know" is fundamental to human judgment, yet AI assistants offer a fluent answer to almost any question. In five experiments (N = 3,132; four preregistered, one direct replication), participants answered difficult questions and could always decline to respond. We engineered the questions so that AI advice was wrong, separating AI use from its accuracy. Merely having access to AI nearly eliminated participants' willingness to suspend judgment, and this held whether the advice was actively requested or simply displayed. Consequently, participants answered more questions but were correct about a third as often as when AI was unavailable—yet their confidence nearly doubled. Incentivizing accuracy and penalizing inaccuracy led participants to seek and follow AI advice less, answer more accurately, and suspend judgment more often, though still far less than when AI was unavailable. As AI suggestions grow ubiquitous and unsolicited, they may not simply affect answer accuracy; they may even alter the metacognitive threshold at which people decide whether they know enough to answer.

The impact of generative artificial intelligence on socioeconomic inequalities and policy making

Valerio Capraro ✉, Austin Lentsch, Daron Acemoglu, Selin Akgun, Aisel Akhmedova, Ennio Bilancini, Jean-François Bonnefon, Pablo Brañas-Garza, Luigi Butera, Karen M Douglas ... Show more

Author Notes

PNAS Nexus, Volume 3, Issue 6, June 2024, pgae191,
https://doi.org/10.1093/pnasnexus/pgae191
Published: 11 June 2024 **Article history ▾**

Abstract

Generative artificial intelligence (AI) has the potential to both exacerbate and ameliorate existing socioeconomic inequalities. In this article, we provide a state-of-the-art interdisciplinary overview of the potential impacts of generative AI on (mis)information and three information-intensive domains: work, education, and healthcare. Our goal is to highlight how generative AI could worsen existing inequalities while illuminating how AI may help mitigate pervasive social problems. In the *information* domain, generative AI can democratize content creation and access but may dramatically expand the production and proliferation of misinformation. In the *workplace*, it can boost productivity and create new jobs, but the benefits will likely be distributed unevenly. In *education*, it offers personalized learning, but may widen the digital divide. In *healthcare*, it might improve diagnostics and accessibility, but could deepen pre-existing inequalities.

Cash et al., 2026. *TICS*

Is AI making us stupid?

Trent N. Cash [1,*], Megan O. Kelly [2], Brooke N. Macnamara [3], and Evan F. Risko [1]

Popular Latest Newsletters *The Atlantic*

July/August 2008 Issue

Is Google Making Us Stupid?

What the Internet is doing to our brains

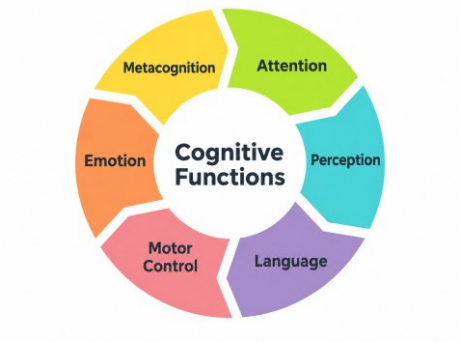
By Nicholas Carr

Skills

- Learned behavior
- Domain-specific
- Aim at accomplishing a task effectively
- Acquired/improved with practice

- Vulnerable to offloading

Cognitive capacities



- Foundational
- Domain-general capacities
- Underlies acquisition and execution of skills

- Resilient to change

Knowledge

- Information storage & retrieval from memory
- Learned content
- Mobilized by skills

- Vulnerable to offloading

- Deskilling
- Mis-skilling
- Never-skilling

AI-uses may induce skilling issues via metacognition

Ke et al., 2026. *Nature Medicine*
Cash et al., 2026. *TICS*

Will AI make us stupid? It is far too early to say with certainty what the long-term effects of offloading cognitive work to AI will be,

Medical students/ early trainees

Trigger

Excessive AI substitution

AI answer-delivery mode during formative training, removes productive cognitive struggle.

Aid

Complement

Mechanisms

Competency acquisition failure

Schema formation requires effortful processing in working memory. AI-supplied answers prevent construction of clinical reasoning architecture.

Calibration paradox

Calibration develops through cycles of independent problem-solving, error recognition and feedback. AI reliance from the outset interrupts these cycles, producing automation bias without technical foundation.

Metacognitive erosion

Sustained AI reliance may reshape professional identity. Trainees may come to understand themselves as intermediaries relaying AI outputs rather than autonomous clinical reasoners.

Outcome and impacts

Never-skilling

Failure to develop foundational clinical competencies.

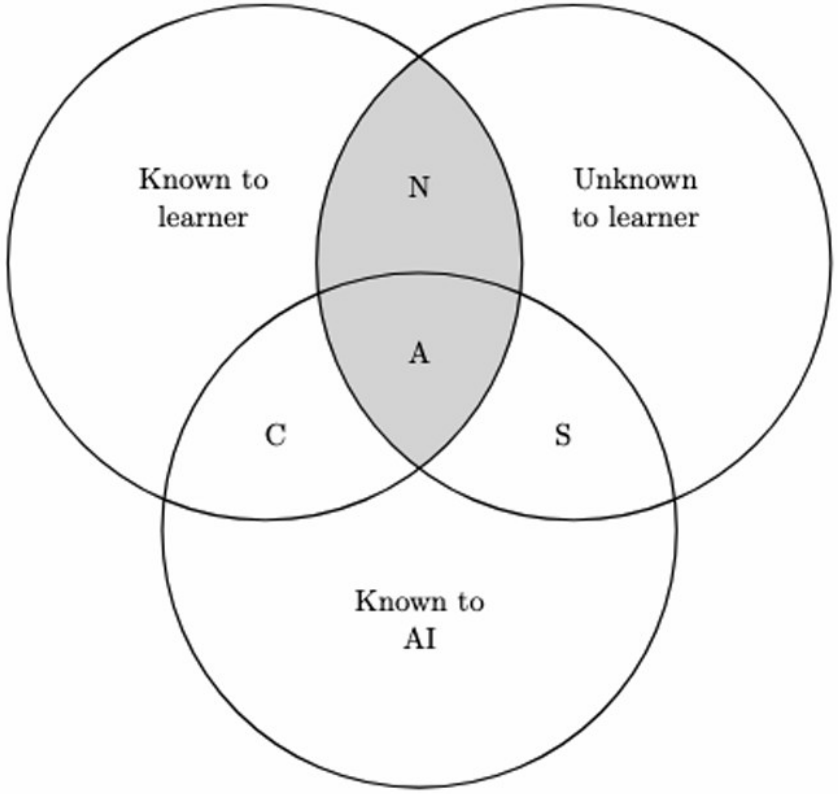
Consequences may impact

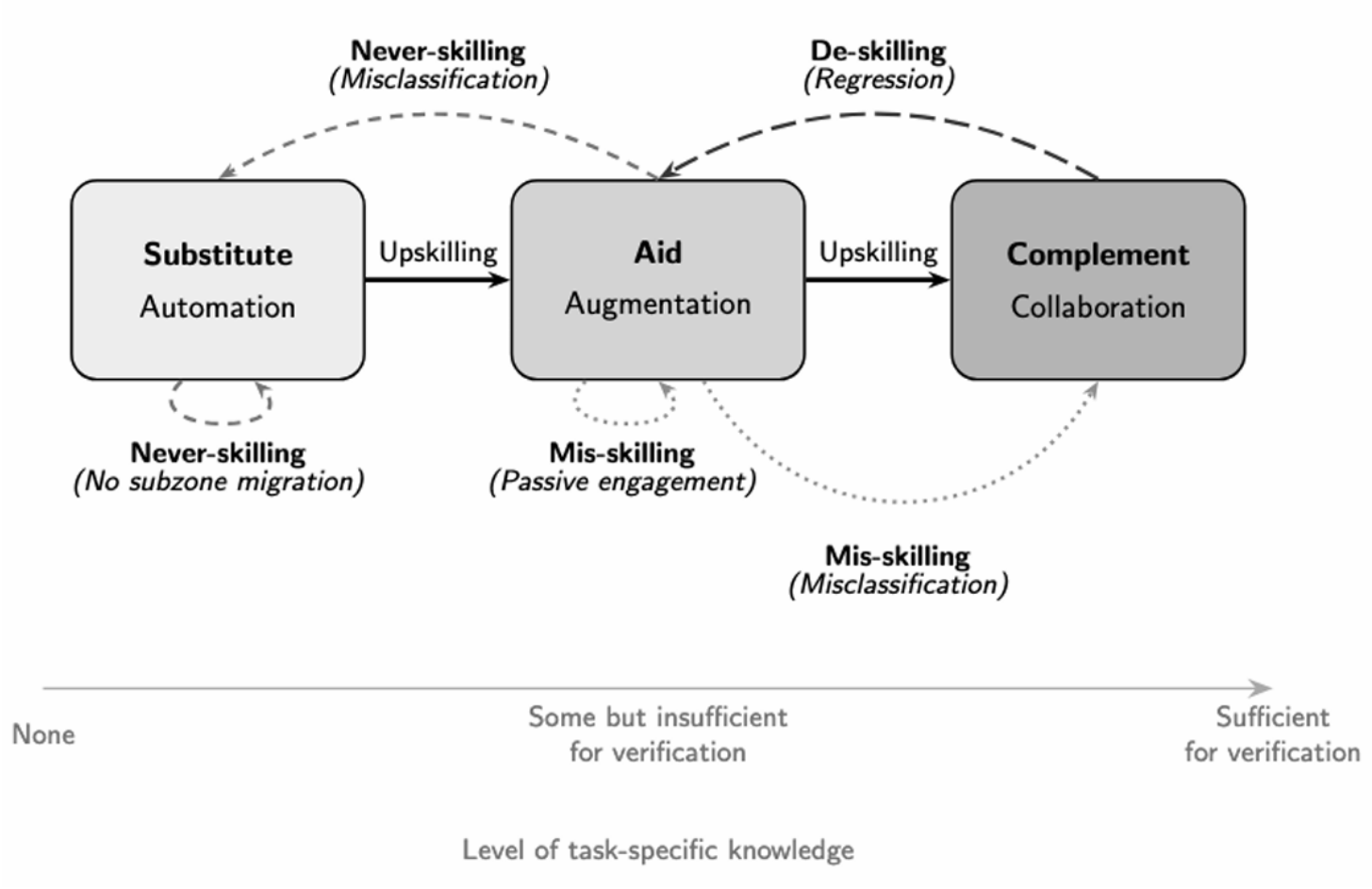
- Patient safety
- Licensing integrity
- Health equity
- Workforce resilience

Model of mechanistic pathway from AI substitution to clinical competency risk during medical training

AI-uses may induce learning acceleration

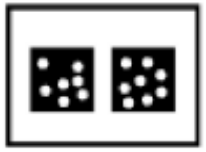
Tsim & Gutoreva, 2026, SCAN
Tsim & Gutoreva, 2026, AI as Teammate





Cognitive & metacognitive redistribution

Without AI



First stimulus

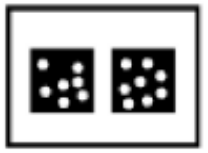
World state (d)

Decision variable (X_{act})

Action (a)

Conf = $P(correct|X_{act})$

With AI



First stimulus

World state(d)

Human processing

Human decision variable (X^H_{act})

Conf = P(correct|X^H_{act})

AI processing

AI decision variable (X^{AI}_{act})

Conf = P(correct|X^{AI}_{act})

Integration / control
(trust, verify, override or delegate)

Decision variable (X_{act})
(final estimate used for action)

Action (a)

Conf = $P(correct|X_{act})$

Judgment's higher-order cognitive operations

1. Problem / inference selection
2. Evidence selection / transformation
3. First-order estimation
4. Second-order estimation
5. Metacognitive control

Each can be:
Human
vs
AI
vs
Joint

Earlier technologies offloaded targeted skills

AI is changing the process by which humans form judgments by breaking/redistributing the higher-order (metacognitive) operations of judgment

Acknowledgments

ISC Marc Jeannerod

cnrs

GATE Lyon / St-Etienne

Social Learning & Decisions SLD Lab

DEPARTMENT OF PSYCHOLOGY

ARTIFICIAL INTELLIGENCE INTERDISCIPLINARY INSTITUTE AT MARYLAND

PROGRAM IN NEUROSCIENCE & COGNITIVE SCIENCE

Julien Benistant

Jean-Claude Dreher

Marie Claire Villeval

CLLE COGNITION LANGUES LANGAGE ERGONOMIE (UMR 5263)

UNIVERSITÉ TOULOUSE Jean Jaurès

Caroline Juliette Charpentier

Rémi Philippe

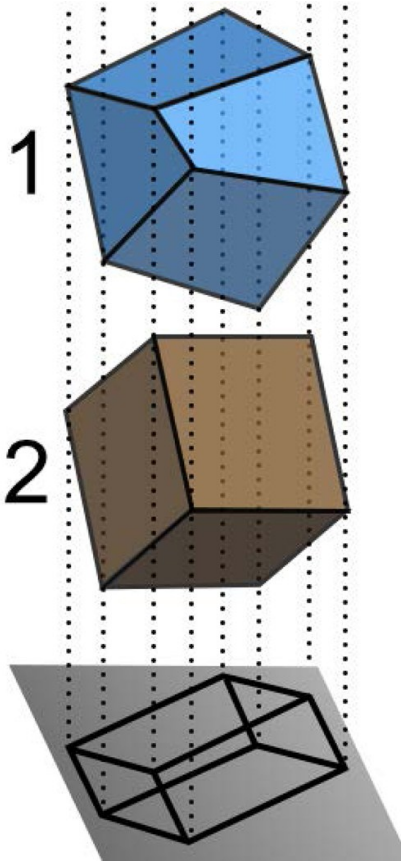
Lucille Geay

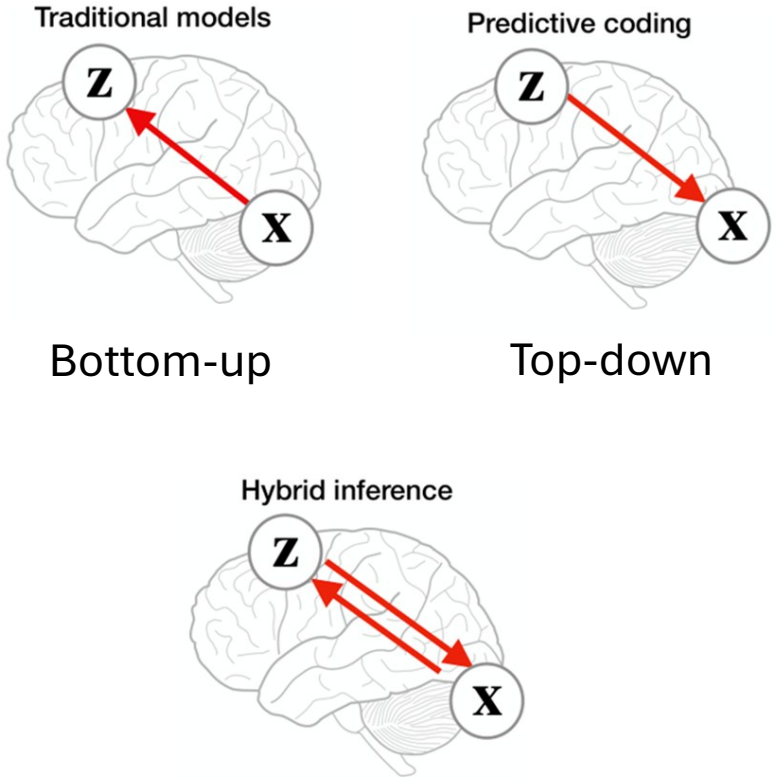
Supmat

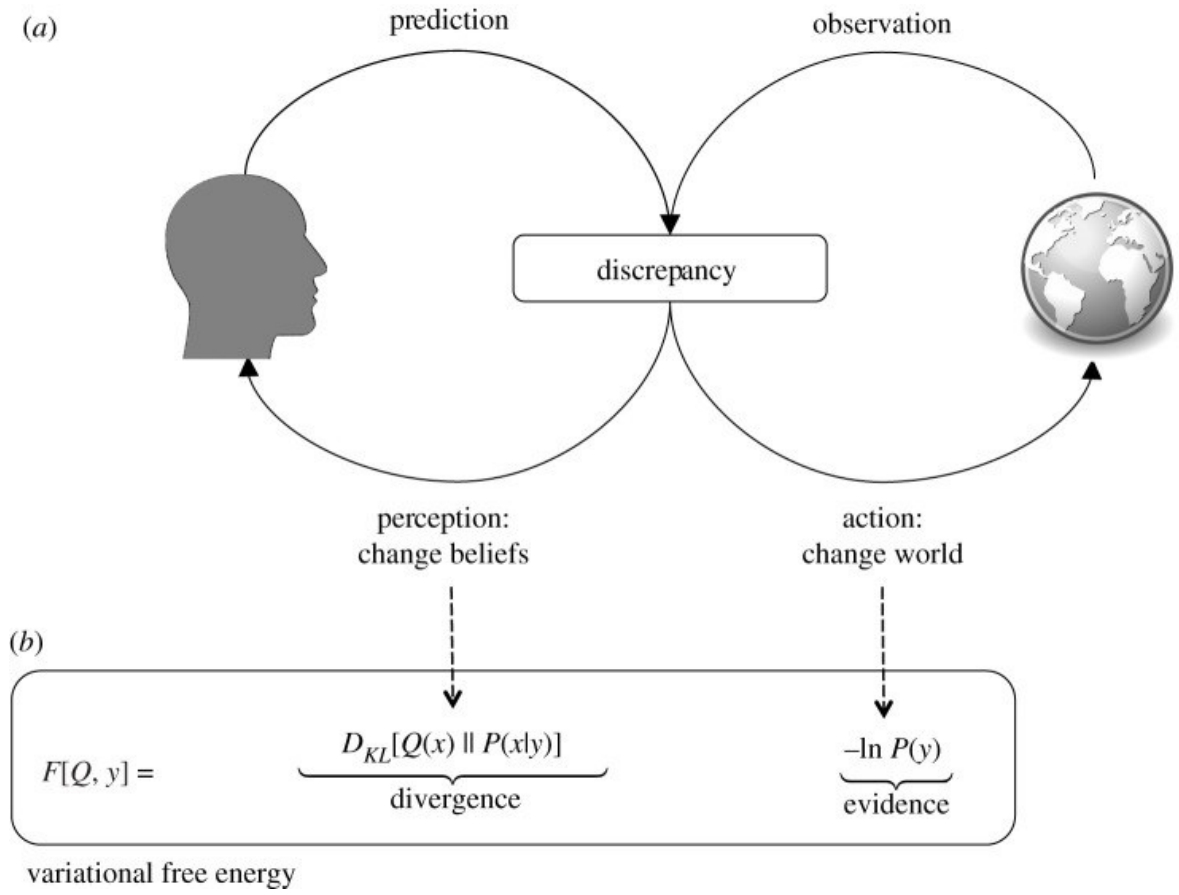
Beliefs-based processing

Born & Bencomo, 2022. *Brain Behav Evol*
Tscshantz et al., 2023. *PLOS Comp. Bio.*
Pezzulo, Parr & Friston, 2022. *Philos Trans R Soc Lond B Biol Sci*

Ambiguity of sensory information: two very different 3D objects can produce identical 2D projections onto the retina

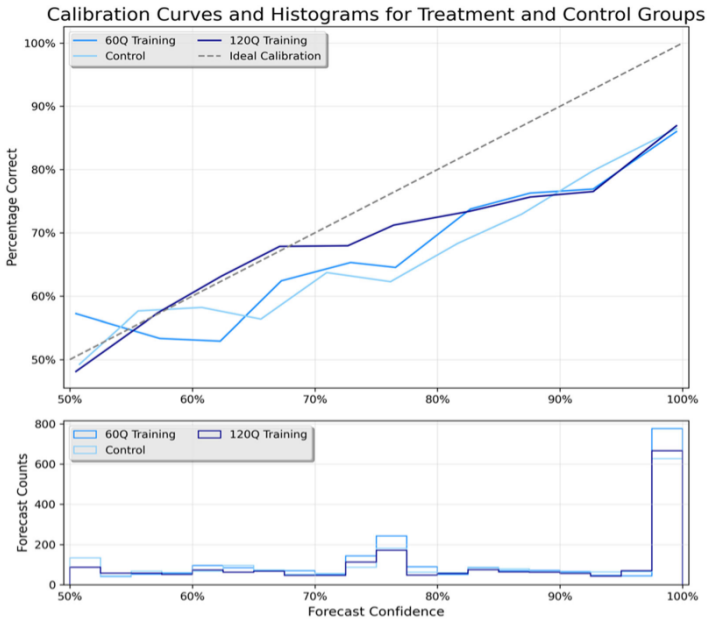




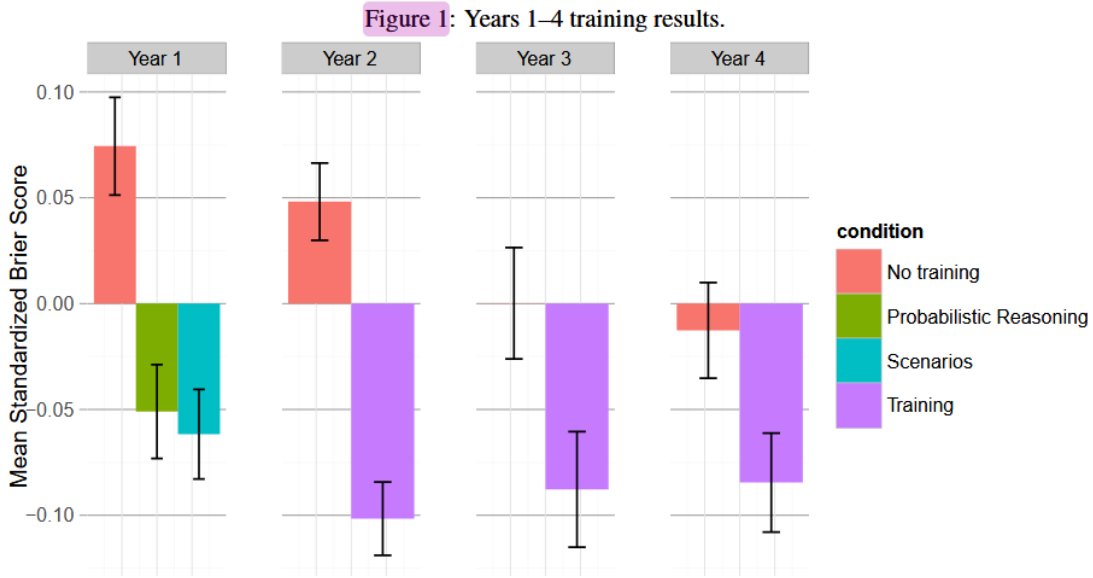


Perception involves inference over possible external causes

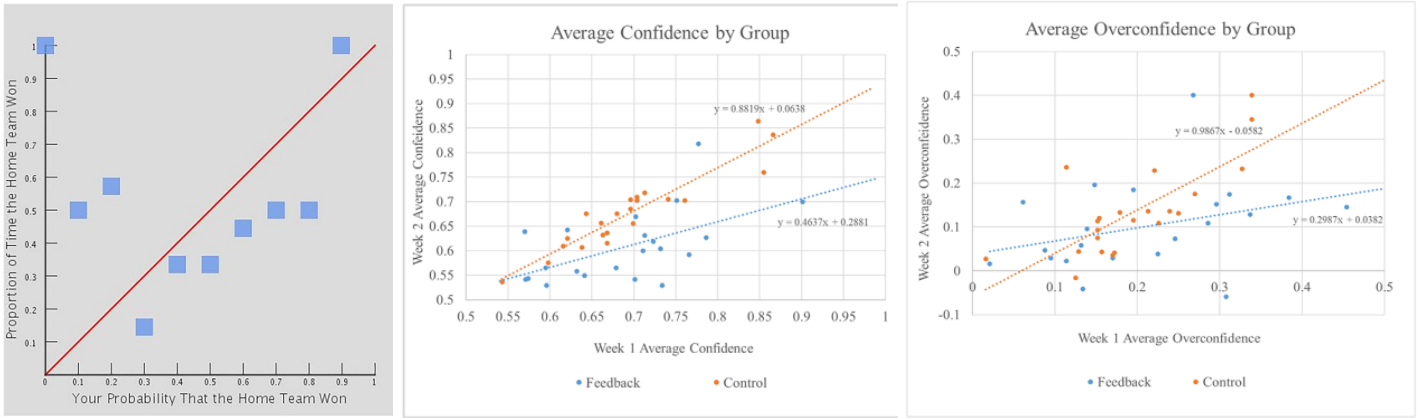
Training probabilistic reasoning and calibration



Gruetzemacher et al., 2024 – students majoring in business and sport culture: prediction of football game winners



Chang et al, 2016 – Geopolitical forecasting
Recruitment: professionals, researchers, alumni associations, blogs, etc.
Training: reasoning principles and probabilistic reasoning



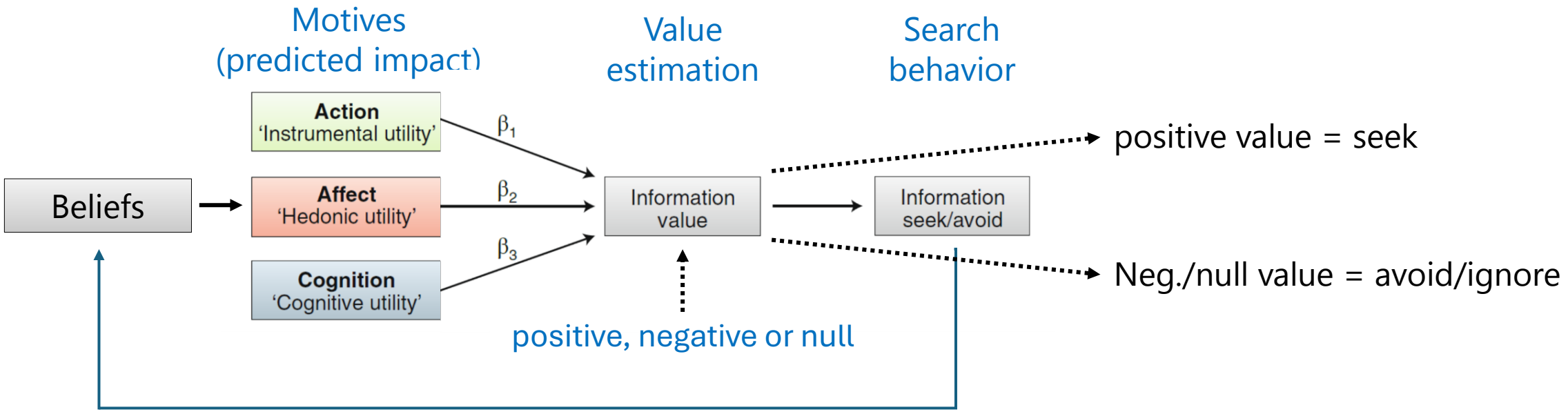
Stone et al., 2023 – left: example of feedback; right: calibration curve
Recruitment: students interested in baseball

Information-seeking motives

Information-seeking motives

Beliefs

Information-seeking motives

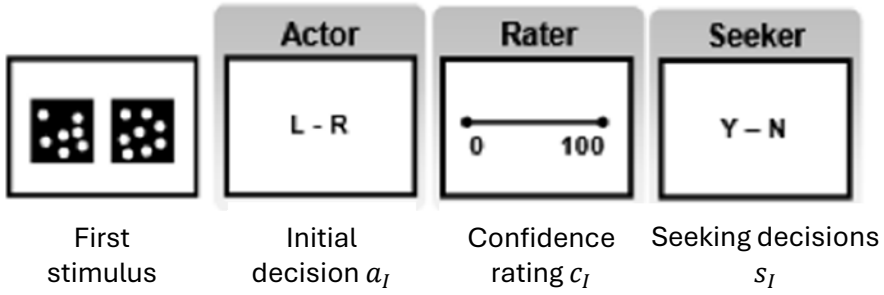


People seek or avoid information depending on:
the value of information & the uncertainties they wish to clarify

(Charpentier et al., 2018. *PNAS*; Shalvi et al., 2019. *Tinbergen Institute Discussion paper*)

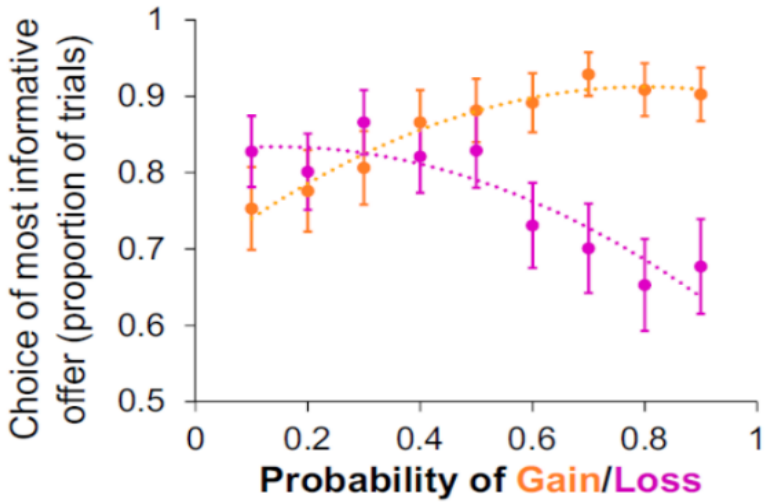
adapted from Sharot & Sunstein, 2020. *Nature Human Behaviour*

Motivated beliefs and information selection

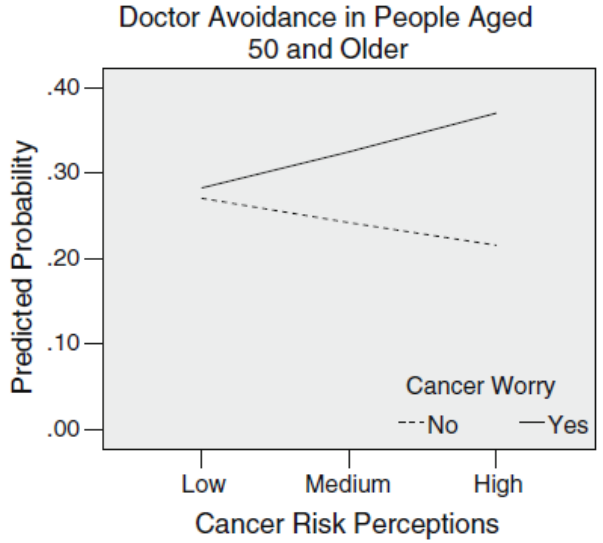


First stimulus | Initial decision a_I | Confidence rating c_I | Seeking decisions s_I

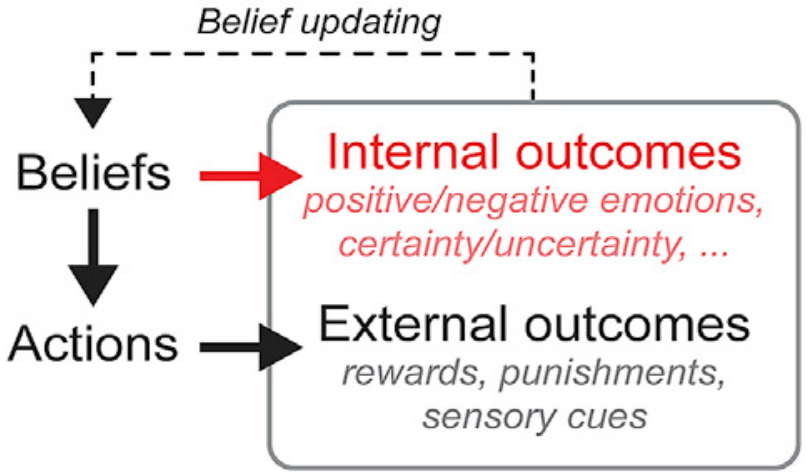
Instrumental utility



Hedonic and cognitive utility



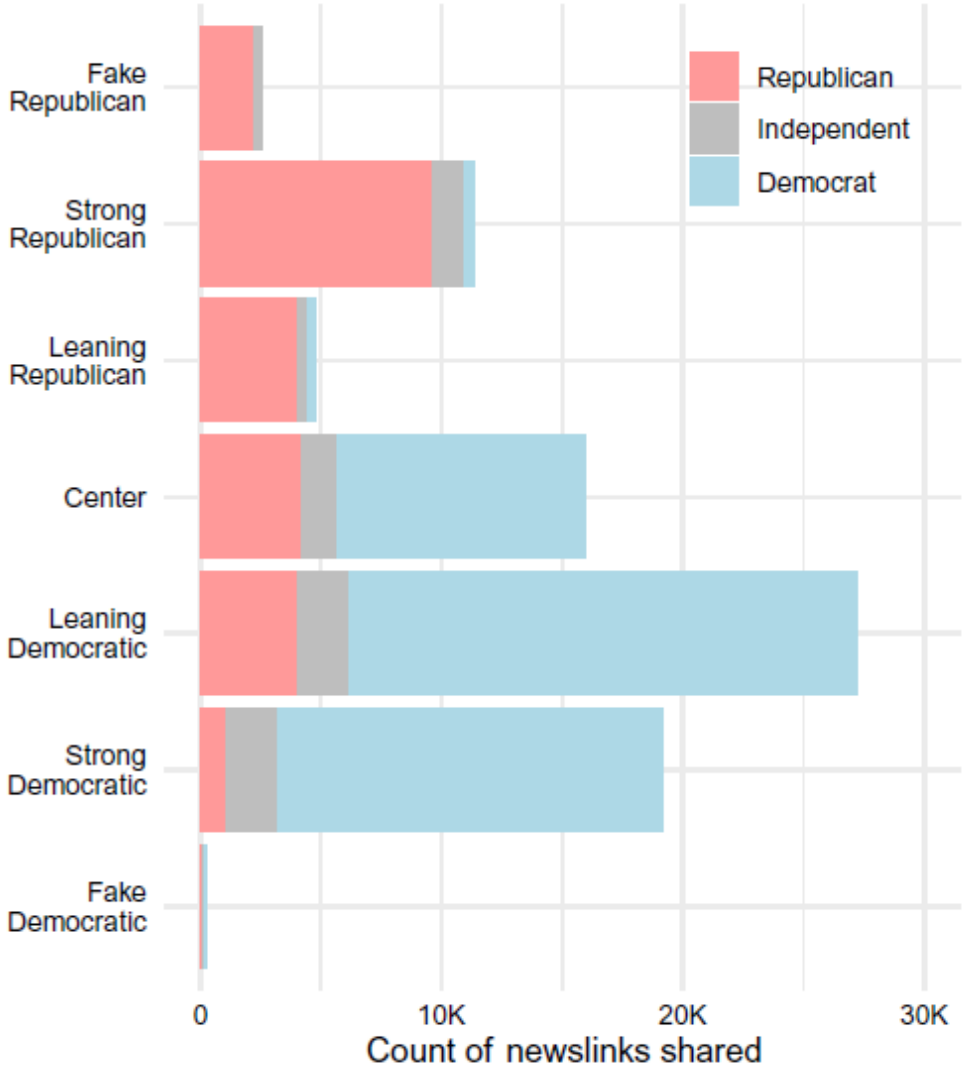
Hedonic utility



Persoskie et al., 2014 ; Charpentier et al., 2018 ; Bromberg-Martin & Sharot, 2020

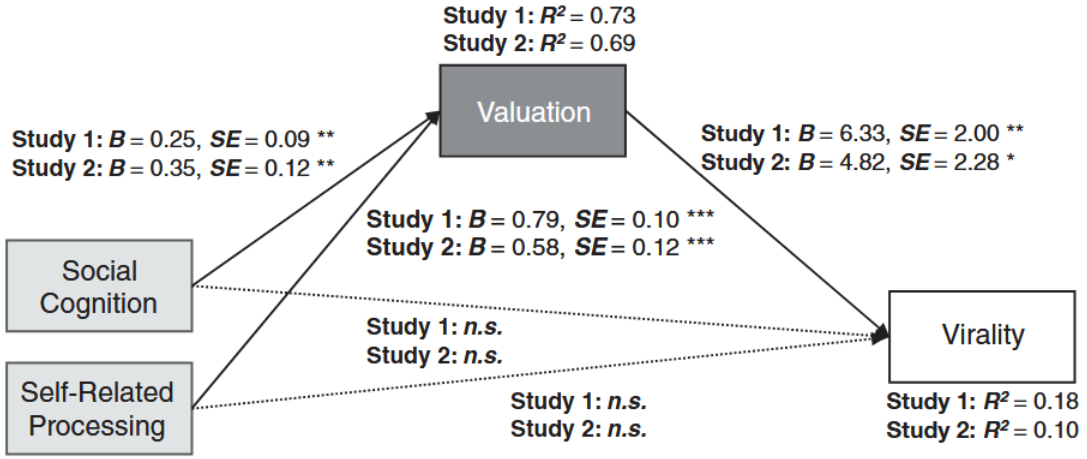
Information-sharing motives

See also Vellani, Glickman & Sharot, 2024,
Three diverse motives for information sharing



Osmundsen et al., 2021. *American political Science Review*

- Veracity-based motivations
- Goal-based motivations (pos./neg.)

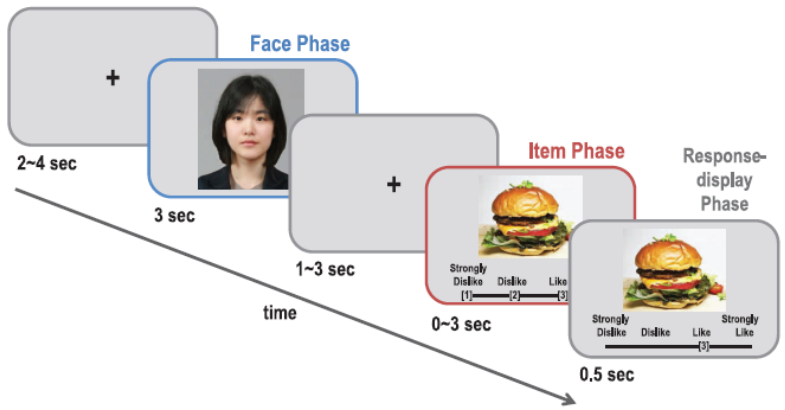


Independent Variable	Study 1 Indirect Effects	Study 2 Indirect Effects
Social Cognition	*B* = 1.59, CI [0.56 – 3.33]	*B* = 1.67, CI [0.24 – 4.87]
Self-Related Processing	*B* = 4.97, CI [2.47 - 8.15]	*B* = 2.81, CI [0.39 – 6.80]

Scholz et al., 2017. *PNAS*; Chan, Scholz et al., 2023. *PNAS*

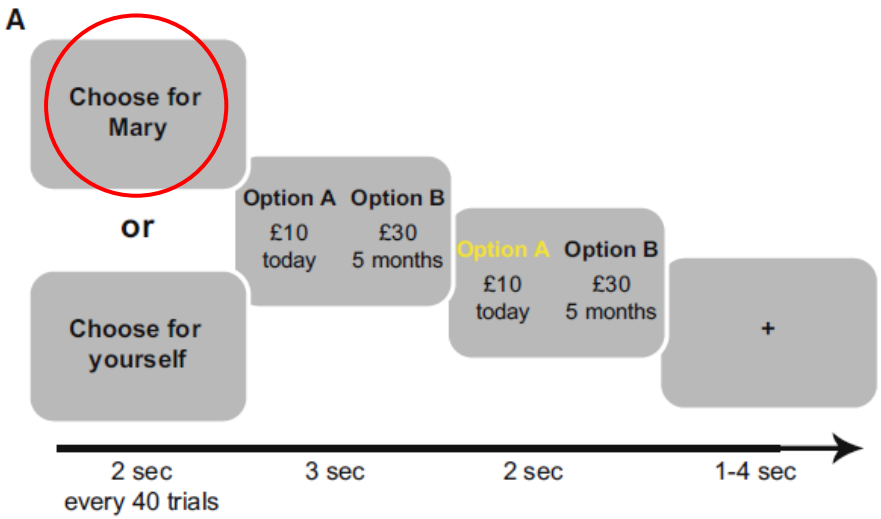
Others'-regarding choices and preferences

Predicting others' preferences from cue



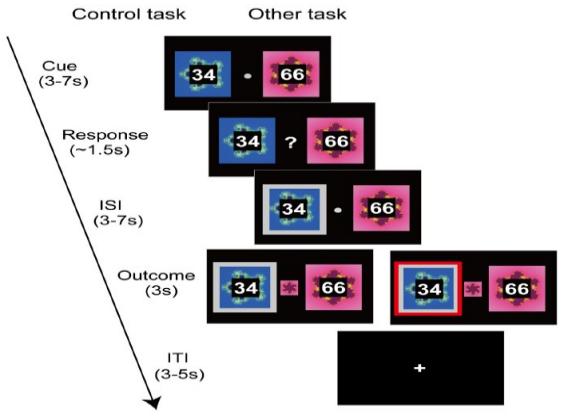
(Kang et al., 2013. *Frontiers in Human Neuroscience*)

Choice on behalf



(Nicolle et al., 2012. *Neuron*)

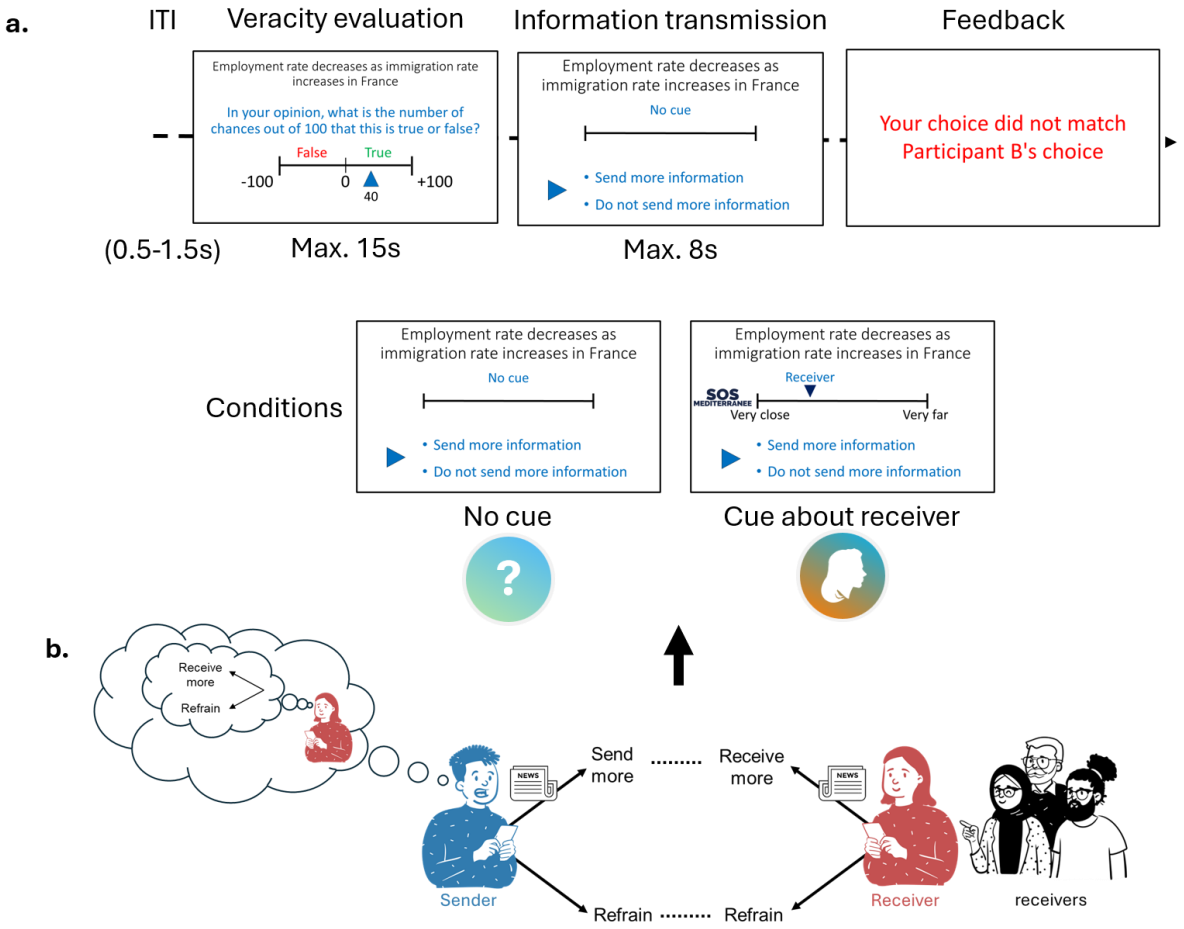
Predicting others' decisions

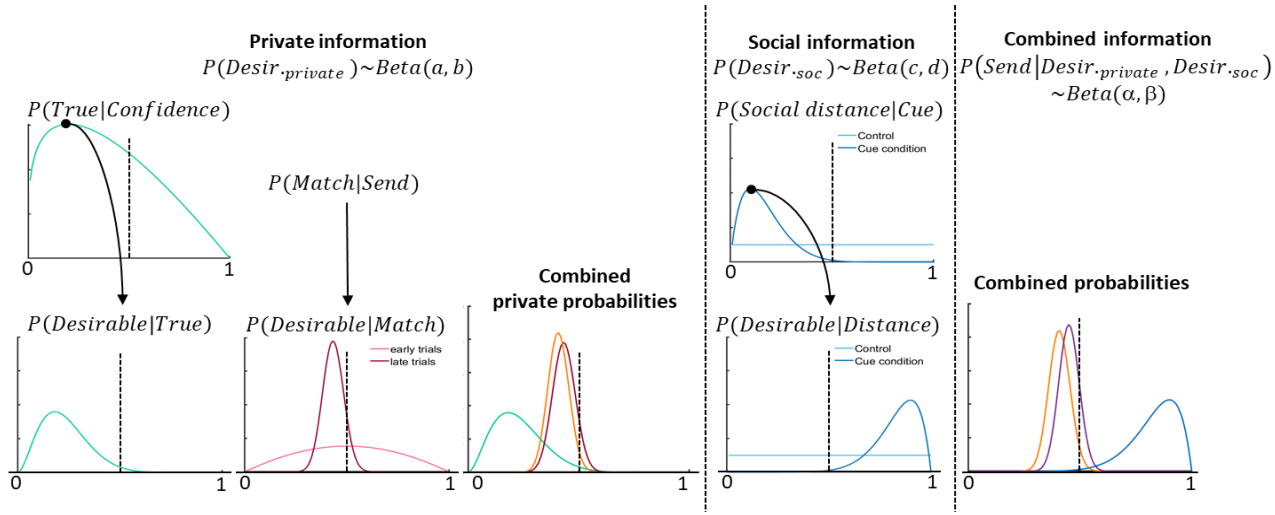


(Suzuki et al., 2012. *Neuron*)

Own valuation + Simulated others' value/action

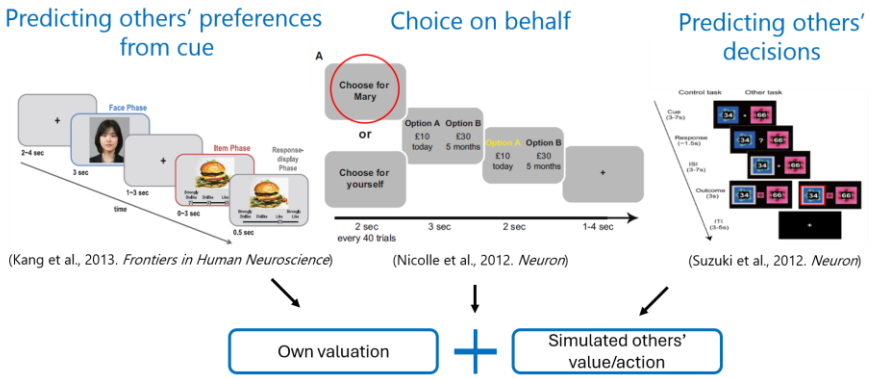
Inferring others' preferences for information



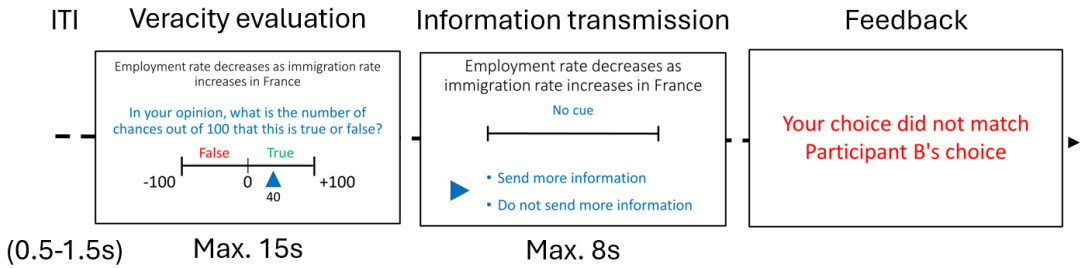


Guigon, Benistant, Villeval & Dreher, *Under review*

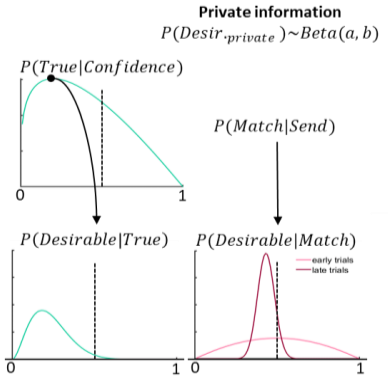
Inferring others' preferences for information

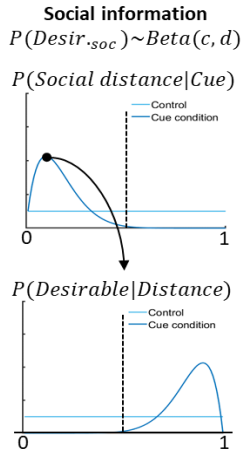


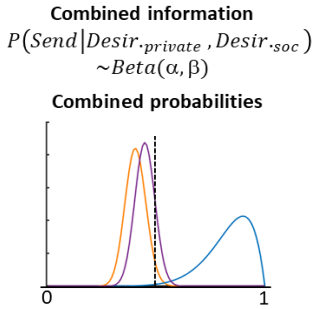
a.

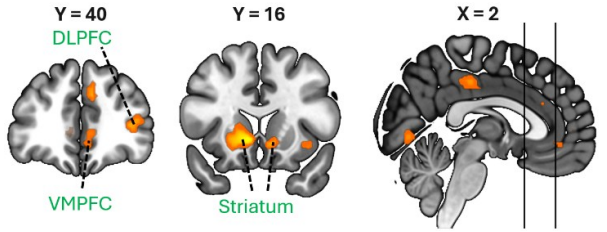


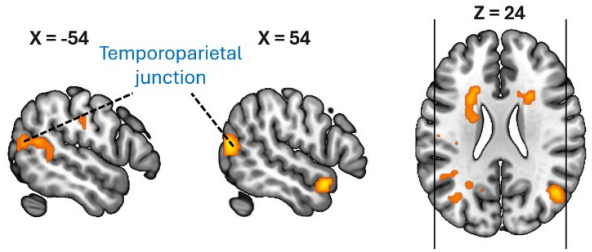
Guigon, Benistant, Villeval & Dreher, *Under review*

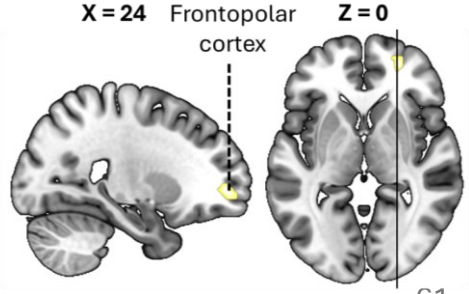




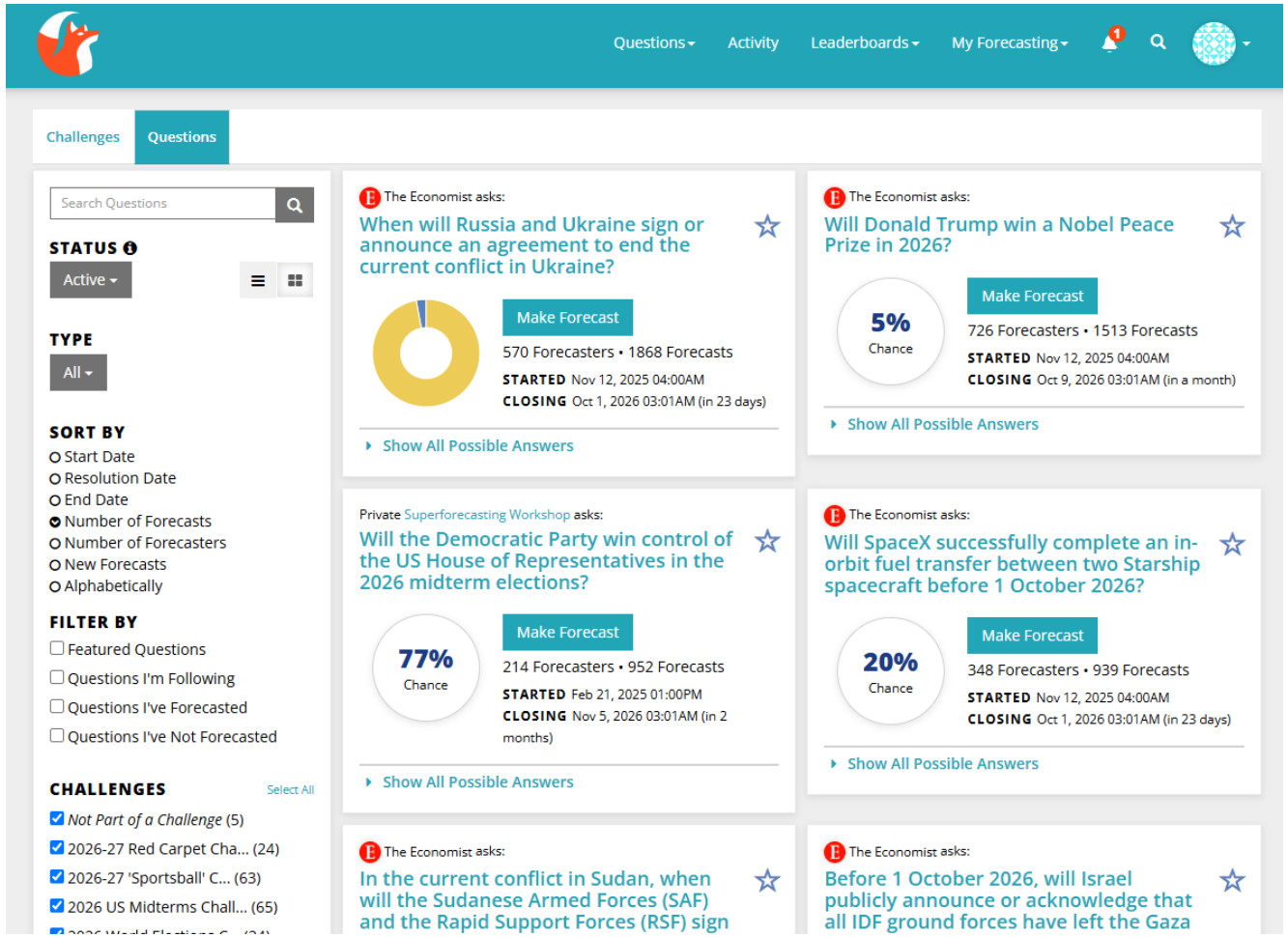




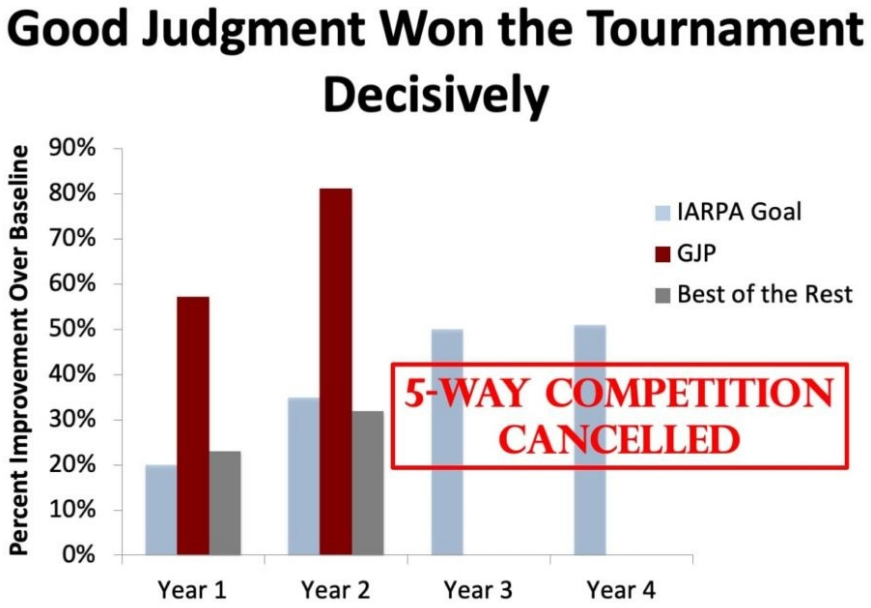




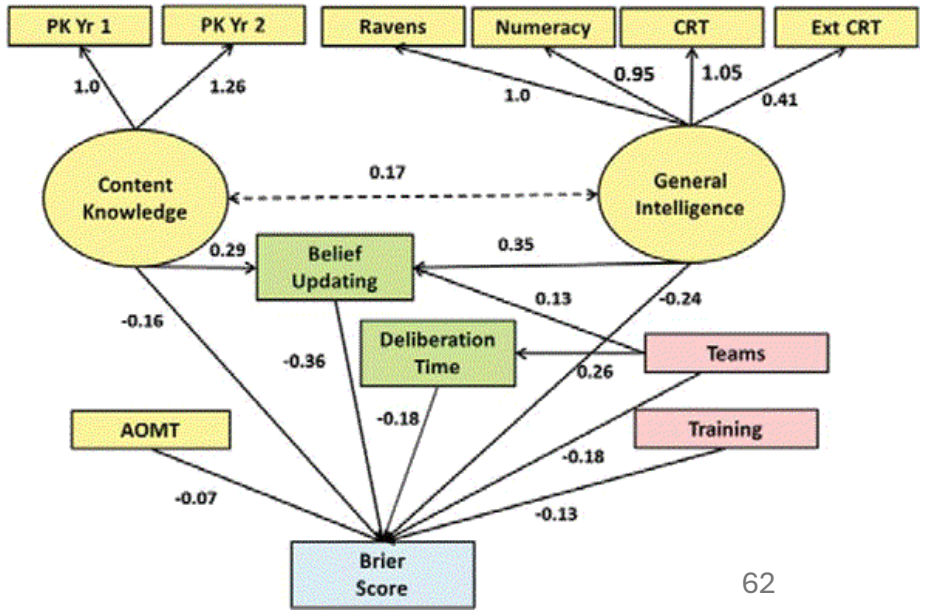
(Super) forecasting



Good Judgment Open (gjopen.com)



See Mellers et al., 2014. *Psychological Science,*
Mellers et al., 2015. *Journal of Experimental Psychology: Applied*



Cognitive operation	Human	AI	Joint
Problem / inference selection	What should I estimate?	AI proposes what to ask / predict	Human selects among AI-proposed problems
Evidence selection / transformation	What information do I attend to / seek?	AI searchers, filters, ranks, summarizes	Human directs and checks retrieval
First-order estimation	What is likely to be X?	AI estimates X	Human + AI estimates combined
Second-order estimation	How likely is the judgment to be correct?	AI supplies uncertainty / calibration information	Human confidence incorporates confidence about AI assessment
Metacognitive control	Seek more? Reconsider? Act?	AI can trigger further search / checks	Human decides when to trust, verify, override, delagate

PASTPRESENTFUTURE

The Future Belongs To Four-Year-Olds

AI is about to make top-tier forecasts available to all. What comes next?

DAN GARDNER

FEB 15, 2026

> This will be a world in which answers are abundant for all. When supply rises, cost falls. When supply is effectively limitless, cost is effectively zero. Answers are going to be cheap, cheap, cheap.
>
> But there are no answers without questions, and AI only answers the questions you pose. Thus, the supply of questions won't change even as the cost of answers plunges.