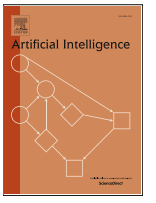
Artificial Intelligence 353 (2026) 104491

Contents lists available at ScienceDirect

Artificial Intelligence

journal homepage: www.elsevier.com/locate/artint



Predictable artificial intelligence

Lexin Zhou [a,b], Pablo A. M. Casares [c], Fernando Martínez-Plumed [a], John Burden [d], Ryan Burnell [e], Lucy Cheke [d,f], Cèsar Ferri [a,g], Alexandru Marcoci [h], Behzad Mehrbakhsh [a,g], Yael Moros-Daval [a], Seán Ó. hÉigeartaigh [d,h], Danaja Rutar [d], Wout Schellaert [a], Konstantinos Voudouris [d,f,i,j], José Hernández-Orallo [a,d,g,*]

[a] Valencian Research Institute of Artificial Intelligence, Universitat Politècnica de València, Spain
[b] Department of Computer Science and Technology, University of Cambridge, United Kingdom
[c] FAR.ai, USA
[d] Leverhulme Centre for the Future of Intelligence, University of Cambridge, United Kingdom
[e] The Alan Turing Institute, United Kingdom
[f] Department of Psychology, University of Cambridge, United Kingdom
[g] Valencian Graduate School and Research Network on AI (ValGRAI), Spain
[h] Centre for the Study of Existential Risk, University of Cambridge, United Kingdom
[i] Department of History and Philosophy of Science, University of Cambridge, United Kingdom
[j] Human-Centered AI, Helmholtz Munich, Germany

ARTICLE INFO

Keywords:
Predictable AI
AI Evaluation
Predictability
Operating Condition
Aleatoric Uncertainty
Epistemic Uncertainty

ABSTRACT

Many areas of artificial intelligence, and machine learning in particular, aim at being probably correct, i.e., valid on average, rather than pursuing the idealistic goal of being provably valid for all inputs. However, AI systems could still be *predictably* valid, such as an imperfect robot deliverer for which we can reliably and precisely predict the task instances for which it is correct and safe, its valid operating range. "Predictable AI" is a nascent research area that explores ways of *anticipating* key validity indicators (e.g., performance, safety) of present and future AI ecosystems. We argue that achieving predictability is crucial for fostering trust, liability, control, alignment and safety of AI, and thus should be prioritised over performance. We formally characterise predictability, explore its most relevant components, illustrate what can be predicted, describe alternative candidates for predictors, as well as the trade-offs between maximising validity and predictability. To illustrate these concepts, we bring an array of illustrative examples covering diverse ecosystem configurations. "Predictable AI" is related to other areas of technical and non-technical AI research, but have distinctive questions, hypotheses, techniques and challenges. This paper aims to elucidate them, calls for identifying paths towards a landscape of predictably valid AI systems and outlines the potential impact of this emergent field.

* Corresponding author.
E-mail addresses: lexinzhouds@gmail.com (L. Zhou), pabloamo@ucm.es (P.A.M. Casares), fmartinez@dsic.upv.es (F. Martínez-Plumed), jjb205@cam.ac.uk (J. Burden), ryan.burnell2@gmail.com (R. Burnell), lgc23@cam.ac.uk (L. Cheke), cferri@dsic.upv.es (C. Ferri), am3159@cam.ac.uk (A. Marcoci), behzaadmehrbakhsh@gmail.com (B. Mehrbakhsh), ymordav@inf.upv.es (Y. Moros-Daval), so348@cam.ac.uk (S. Ó. hÉigeartaigh), rutar.danaja@gmail.com (D. Rutar), wschell@vrain.upv.es (W. Schellaert), kv301@cam.ac.uk (K. Voudouris), jorallo@upv.es (J. Hernández-Orallo).

https://doi.org/10.1016/j.artint.2026.104491
Received 2 October 2023; Received in revised form 7 January 2026; Accepted 26 January 2026
Available online 30 January 2026

1. Why predicting validity is central to AI

AI is set to transform every aspect of society, yet this progress has brought a validity problem: it is becoming increasingly hard to anticipate exactly when and where an AI system will give a valid result or not. Consider a delivery robot that fails to transport a parcel to its destination. The key question is not whether we can explain the failure *ex-post*, but rather whether we could have anticipated it *ex-ante*. If we cannot anticipate when and where AI systems can be deployed effectively and safely, then we are at the mercy of a *lottery* of generalisation issues, adversarial examples and instruction ambiguities [1]. For instance, image classifiers suffer *Clever Hans* effects [2,3], agents exhibit unanticipated reward hacking phenomena [4,5], and language models display unexpected emergent capabilities [6,7] , hallucinations [8,9] or other hazards [10,11].

General-purpose AI models, in particular, are drawing attention to several other long-standing problems in AI. First, we do not have a specification against which to verify these systems; there is no single task or distribution for which to maximise performance (and maximising performance on proxies is insufficient [12–14]). Second, we do not expect the AI system to work well for every input; depending on the context, there might be value if it just works for some inputs [15,16], e.g., self-driving cars should be deployed on conditions under which they are predictably safe [17–19]. Third, mechanistically anticipating every single step, text or video generated by AI is impractical, and might even be an unnecessary or undesirable objective; we also want AI systems to generate outputs that we cannot generate ourselves, especially those that require considerable effort or are meant to be creative [20].

As a concrete example, Fig. 1 represents six figurative AI systems (A, B, C, D, E and F) commanding self-driving cars. Although they all have the same expected validity (performance of 62.5%), the distribution of this performance varies according to *windingness* and *fogginess*: certain systems (in particular A and B, but also C and D) are more easily predictable than others, given the two known features.

In this paper, we argue that all else being equal, more predictable AI systems are preferable. In Fig. 1, we observe that, with a simple univariate logistic function on the feature windingness, we can easily build a predictor of A's performance, $\hat{p}_A$, that can model p_A quite well. Similarly for B using fogginess. With bivariate functions we can model C and D quite well too, but E and F seem to require more complex function families to capture the patterns of validity, if they exist at all. This example illustrates that the goal of AI and its evaluation may not be building *provably correct*, i.e., infallible systems, not even *probably correct* AI systems that ensure a higher degree of performance on average (e.g., a self-driving car that is 99.9% valid), but building AI systems that are *predictably valid*, such that we can anticipate, precisely and reliably, where this validity is found.

Pursuing more predictable AI is pivotal because current AI systems and societal AI futures are largely unpredictable for humans [20, 22,23], especially when compared to predicting human performance [24], and therefore it is difficult to guarantee beneficial system development and deployment. Following with the self-driving car example, achieving predictability is an essential precondition for fulfilling key desiderata of AI, such as trust, liability, control, alignment and safety:

- **Trust** in AI "is viewed as a set of specific beliefs dealing with [validity] (benevolence, competence, integrity, reliability) and *predictability*" [25,26]. Self-driving car users will trust their car not because they understand the complex underlying AI system driving it, but because they can use the weather and road conditions to predict whether the car can operate successfully and safely. However, the right level of trust between overreliance and underreliance [11] is rarely met since "the unpredictability

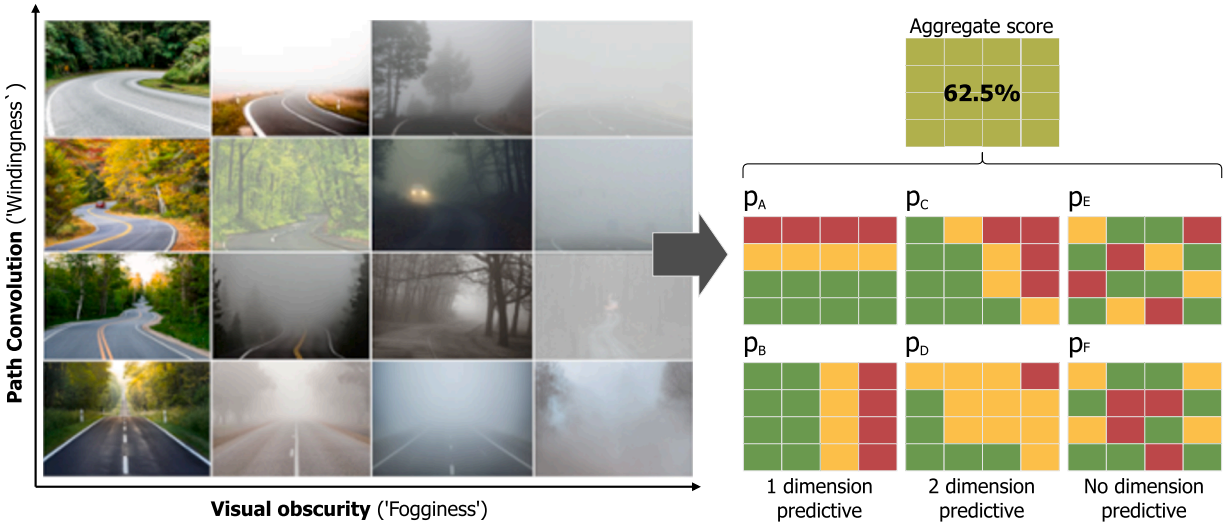


Fig. 1. A driving scenario where six AI systems (A,B,C,D,E,F) control self-driving cars, with all systems having the same expected validity (62.5%); the grids of colour green, yellow and red represent fully valid, partially valid, and invalid regions, respectively. The distributions of validity for the six systems, $p_A, p_B, p_C, p_D, p_E, p_F$, differ across windingness and fogginess. Which scenarios are easily predictable with those dimensions? Illustration adapted from [21].

of AI mistakes warrants caution against overreliance" [27]. Predictability is an essential property of AI that enables reliable assumptions by stakeholders about the outcome [28].

- **Liability** for AI-induced damages applies when an alternative decision could have avoided or mitigated a *foreseeable* risk. Does an accident happen in a situation that could have been anticipated, such as a foggy trip in the mountains? Currently, "AI unpredictability [poses] significant challenges to proving causality in liability regimes [29]. The question is then to determine if harm was *predictable*, not by the system or by its designers, but by any available and conceivable diligent method.
- **Control** of AI refers to being able to stop a system, reject its decisions and correct its behaviour at any time, to keep it inside the operating conditions. Keeping a human on the wheel that can take control for difficult conditions can even make things worse if these conditions cannot be reliably determined by the human driver with sufficient anticipation. Control requires effective *oversight*, not an uninformed human in the loop giving a "false sense of security" [27,30,31], as "*predictability* is a prerequisite for effective human control of artificial intelligence" [32].
- **Alignment** of AI has multiple interpretations focusing on the extent to which AI pursues human instructions, intentions, preferences, desires, interests or values [33]. When taking the self-driving car after a heavy dinner, what if the user requests the shortest path home, and the car takes a peaceful detour instead of a winding road that would have made the user feel sick, as happened on previous occasions for this user? The *validity predictability problem* must include the human user and the context of use as inputs.
- **Safety** in AI aims to minimise accidents or any other "harmful and *unexpected* results" [34]. The user will accept that the car does not take several routes if they have high risk estimates or high uncertainty in these estimates. One of the key principles of safety is to deploy systems only under operating conditions where they can be *predictably safe*, i.e., low risk of a negative incident. A reliable rejection rule to implement a safety envelope depends on accurately estimating when the probability of harm exceeds a safety threshold [35].

The self-driving car example is representative of the key implications of predicting validity, not only in isolated AI systems, but *AI ecosystems* where AI systems interact with humans and other AI systems (other cars—self-driving or not—, pedestrians, etc.). How can we study AI predictability in general, and frame it in such a way that it takes the central stage it deserves as a research area? This is the goal of this paper.

First, we introduce Predictable AI as a distinct research area, formally characterising the concept of predictability, what it is and what it is not, distinguishing it from other research areas (Section 2). Next, we establish a comprehensive framework that unifies the problem space, providing the necessary tools to quantify and reason about predictability (Section 3). Following this, we use seven concrete scenarios to illustrate our framework in action, highlighting a wide range of AI systems and prediction types (Section 4). We survey practical approaches to validity forecasting, including human, self-estimating and external predictors, and evaluate the implications and trade-offs involved in maximising both validity and predictability (Section 5). Finally, the paper identifies major open challenges and highlights promising opportunities for advancing the field (Section 6). The main contribution of this paper is to lay a foundation for systematically studying and improving the predictability of validity in AI, reinterpret several scattered efforts under the same framing to make synergies possible, and construct a new paradigm that changes the priorities of AI design and evaluation.

2. What is predictable AI? What is it not?

AI Predictability is the extent to which key *validity indicators* of present and future AI ecosystems can be anticipated. These validity indicators are measurable outcomes (resulting from interactions between tasks, systems and users) such as performance and safety. AI ecosystems range from single AI systems interacting with individual users for specific task[1] instances, all the way to complex socio-technological environments, with different levels of granularity. At one extreme of the spectrum, predictability may refer to the extent to which any validity indicator can be anticipated in a specific context of use, such as how predictable a single AI system's performance or safety is when completing a user command. For the previous self-driving car example, we want to anticipate the validity for each instance—each journey—beforehand, i.e., predict whether the car will reach the destination safely. At the other end, Predictable AI may refer to the ability to predict where the field of AI is heading, anticipating future capabilities and safety issues of the entire AI landscape several years ahead. For instance, we want to anticipate what capabilities new large language models will have a year ahead, by building scaling models or using some other forecasting techniques. While we consider the whole spectrum in Predictable AI, the rest of the paper will include more examples of the former, specific-context usage, since it represents a more clear paradigm change for AI research that can be translated into immediate benefits.

At first glance it may seem that full predictability is always desirable, yet there are a variety of situations in which it is not necessary or practical to anticipate the ecosystem's full behaviour [20,36–38]. We do not need to—in fact we cannot—predict every single action a self-driving car will make during each journey. We only need to predict whether the car will take us to our destination safely. We want to give freedom to the AI driver to find out a particular set of actions that achieve this goal. After all, the promise of original, unpredictable outputs is one of the motivations for using AI in the first place [39]. This is especially the case for generative AI models, where the novelty of outputs is key, such as generating the text for a wedding invitation or an illustration for a magazine. In these situations, predicting performance, safety, timelines, or some other abstract *validity indicators* makes more sense than predicting

[1] A task is a collection of instances. For example, a task can be defined as a two-number addition problem, while "1 + 1 =" and "712 + 123 =" are two distinct specific instances in the two-number addition task. Similarly, each journey in our self-driving car example is an instance of the task of driving.

Table 1
Qualitative comparison of aleatoric (irreducible) and epistemic (reducible) uncertainty for three prediction targets. Forecasting full environment or AI behaviour is often dominated by aleatoric uncertainty, whereas predicting coarse-grained environment-AI validity can be less demanding (though still not fully free of aleatoric effects).

Kind of Uncertainty	Environment Behaviour	AI Behaviour	Environment-AI Validity
Aleatoric Uncertainty	Unpredictable	Unpredictable	Unpredictable
Epistemic Uncertainty	Very hard to predict	Hard to predict	Potentially predictable

Table 2
Validity and predictability for an AI system (e.g., an LLM) that classifies complaints into one of five departments $\{d_1, \dots, d_5\}$. Validity is accuracy of the 'base' classifier and predictability is accuracy of the 'alt' problem, how good we can get predicting whether the 'base' classifier is correct.

System Category	Behaviour	Expected Validity ($\mathbb{V}$)	Predictability ($\mathbb{P}$)
Constant (d_4)	Always outputs d_4 (with prior $p(d_4) = 0.15$)	0.15	0.70[a]
Random	Uniformly samples a department (1/5 each)	0.20	0.20
Optimal	Always selects the *correct* d_i	1.00	1.00
Pessimal	Always selects an *incorrect* d_i	0.00	1.00

[a] With $\mathcal{F}$ the set of all computable predictors and aleatoric uncertainty $U_a = 0.3$, the ceiling is $\mathbb{P}_{\max} = 1 - U_a = 0.7$.

full behavioural traces. This becomes more patent when we distinguish between aleatoric uncertainty (irreducible randomness) and epistemic uncertainty (uncertainty due to limited information or modelling capacity, which can in principle be reduced) [40]. Table 1 summarises how these components typically arise for three prediction targets. If we attempt to predict full *environment behaviour*, most variation is effectively aleatoric (e.g., during a car journey, the movements of other vehicles and pedestrians are difficult to forecast across relevant time scales). Predicting full *AI behaviour* is similarly challenging: modern systems may be intrinsically stochastic and their actions are tightly coupled to environment dynamics (e.g., chosen trajectory, braking points). In contrast, predicting *environment-AI validity* is a less demanding objective, because we can choose those indicators that are more predictable and for which we can determine the outcome with higher confidence (e.g., whether the car reaches the destination safely). Aleatoric uncertainty remains in all three cases (e.g., a sudden manoeuvre by another car causing an accident), but it is arguably smaller in the Environment-AI Validity case.

In simple cases, such as the self-driving car example, we can define validity V as the random variable that integrates all the utilities and costs we care about, concerning the quality and safety of journeys. A particular AI system s in this scenario can simply be a function that maps an instance i, a journey, to a behaviour (the output or actions of the system), while the validity function maps that journey i to a validity value. As this process is stochastic, we can represent this as a conditional validity probability over instances, $p_s(V \mid i)$. Then, our AI predictability problem consists of constructing an estimator, an *assessor*, for that probability, $\hat{p}_s(V \mid i)$. One possible way of doing this is through machine learning. By learning from examples of the use of the AI system (pairs of instance and validity score) we can build an assessor $\hat{p}_s$ for one AI system s (or a generic assessor parametrised with s). To this purpose we can use any family of techniques $\mathcal{F}$ in machine learning. For the self-driving car example in Fig. 1, the family of linear functions using fogginess and windingness as inputs would be sufficient to build good assessors for self-driving cars p_A, p_B, p_C and p_D, but not for p_E and p_F.

We additionally introduce a simpler classification example that highlights how validity and its predictability can diverge. Table 2 grounds four AI systems for a real-world task in which a classifier based on a large language model (LLM) must route consumer complaints to one in five departments ($\{d_1, \dots, d_5\}$). Empirical priors are $p(d_1) = 0.35$, $p(d_2) = 0.25$, $p(d_3) = 0.20$, $p(d_4) = 0.15$ and $p(d_5) = 0.05$. *Validity* (V) in this case is simply classification into the correct department, while *expected* validity $\mathbb{V}$ is represented by the average accuracy over the instances of the task. *Predictability* ($\mathbb{P}$) is the probability that an *optimal* predictor in a chosen family of predictors $\mathcal{F}$ can correctly anticipate whether the system will be valid on a given task instance. Let us consider $\mathcal{F}$ as the set of all computable predictors. If we assume that the intrinsic aleatoric uncertainty of the task is $U_a = 0.30$ then the theoretical ceiling is therefore $\mathbb{P}_{\max} = 1 - U_a = 0.70$. In these conditions, we can determine $\mathbb{V}$ and $\mathbb{P}$ for several AI systems. The constant AI system that always outputs d_4 attains $\mathbb{V} = 0.15$ and its predictability (by an assessor that reduces all the epistemic uncertainty) is $\mathbb{P} = 0.70$. However, a uniformly random router achieves $\mathbb{V} = \mathbb{P} = 0.20$, showing that random behaviour in an AI system can make predictability much worse than what the intrinsic aleatoric uncertainty of the task determines (0.70), as randomness adds the aleatoric uncertainty (stochasticity) of the AI system. Finally, as shown in Table 2, the optimal and pessimal AI systems remain perfectly predictable ($\mathbb{P} = 1$) despite opposite validity scores, clearly illustrating that validity and predictability are very different things.

The examples of the self-driving car and the department classifier illustrate that predictability of AI systems in their ecosystems will depend on (1) the ecosystems in the first place, usually conceptualised as the problem the AI system tries to solve (the driving problem or the classification problem), (2) the AI systems that are operating in that context (the self-driving car and the classifier, respectively), (3) the validity indicator (safe, effective journeys and assignment to the correct department, respectively) and (4) a family of assessors that may be limited by representational or computational constraints (logistic functions using two features

and all computable predictors, respectively). These four elements determine the capability for predicting validity in a particular AI ecosystem (anticipating which journeys the AI-driven vehicle is going to complete safely, or which complaints the classifier will send to the correct department).

Predictable AI is relevant across a wide spectrum of application areas, each of which poses its own challenges in terms of anticipating system validity. To illustrate the scope and practical significance of validity prediction, Table 3 provides a selection of scenarios where the outcomes of AI ecosystems need to be predicted. These examples differ in many details (e.g., the input features, the number of subject models or task instances, the length of temporal horizon, who predicts and how), which we formalise and further discuss in Section 3. Nonetheless, we must consider what these examples have in common: the need to predict certain validity indicators or outcomes in a context where AI plays a fundamental role. This 'validity prediction' is an 'alternative' ('alt') problem, differing significantly from the original task of the AI system, referred to as the 'base' problem (driving a car to a destination or assigning a complaint to a department). We will characterise this formally in Section 3. We will often refer to the AI system that solves the original task as the 'base system', as opposed to the 'alt system', the assessor that predicts the base system's validity.

From the perspective of systems theory or social sciences, predicting outcomes, and the complexity of this prediction, are expected and natural. Within particular areas of computer science, such as software engineering and machine learning, however, the traditional focus has been set on short-term predictions about individual systems using aggregate statistics, such as time between bugs or average error. This is manifested in predictive testing in software engineering [41,42] and model performance extrapolation in machine learning [43]. Nevertheless, for many AI systems, and especially general-purpose AI systems [44, Art. 3], it is no longer feasible to aim for full verification (provable validity, with all journeys always being successful, as per the previous example) and it is no longer sufficient to have average accuracy extrapolation (probable validity, with an estimated 62.5% of valid journeys, in the previous example). Instead, we need predictable validity, detailed predictions given specific instance-level contextual demands, such as the question asked, order given or the conditions of the tasks, as in Fig. 1. We also need to consider longer-term multiple-system scenarios, more commonly covered in AI forecasting [45,46], such as predicting whether AI will be able to do a particular job in a certain number of years [47–50]. These considerations, among others, are present across the examples in Table 3, which will be detailed in Section 3 in tandem with the formalisation of Predictable AI.

Ensuring that an AI ecosystem is robust and safe across all possible inputs, conditions, users or contexts can be a formidable challenge and may not always be necessary. A more practical goal, instead, is to reliably predict where exactly the ecosystem will resolve favourably or not. Given these validity prediction models, we can consider which pair of base system and alt model (the validity prediction model) gives the best trade-off between maximising validity and predictability of that validity. Identifying and selecting AI ecosystems with predictable validity should be the key focus of the field of Predictable AI, especially in the age of general-purpose AI.

Because Predictable AI is so ingrained in key issues of AI, such as trust, liability, control, alignment and safety, seen in Section 1, it must be closely related to other paradigms and frameworks of analysis that share these goals, such as explainable AI, interpretable AI, safe AI, robust AI, trustworthy AI, causal AI, AI fairness, sustainable AI, responsible AI, etc. Table 4 summarises the most relevant ones and how Predictable AI differs from them.

3. AI ecosystems and predictability

This section presents a unified framework for defining and quantifying predictability in AI. For simplicity, in this section, we will introduce a formalisation that only considers a single AI system, instance and user at a time. A more general definition that considers the time evolution of complex AI ecosystems can be found in Appendix A.

We consider an ecosystem $\mathcal{E}$ as a distribution of instances, systems and users $\langle i, s, u\rangle$. In this ecosystem, validity is defined by $p(v \mid \langle i, s, u\rangle)$. The expected validity $\mathbb{V}$ is defined as:

$$\mathbb{V}(p, \mathcal{E}) := \mathop{\mathbb{E}}_{\langle i,s,u\rangle\sim\mathcal{E}}[V] = \mathop{\mathbb{E}}_{\langle i,s,u\rangle\sim\mathcal{E}}\left[\int_v v \cdot p(v \mid \langle i, s, u\rangle) dv\right] \tag{1}$$

The traditional goal of AI has been to build AI systems and deploy them such that they maximise $\mathbb{V}$. For instance, in the example in Fig. 1, the expected validity is 62.5% for a given distribution of ecosystems $\mathcal{E}$ (conditions, other cars, people, etc.) when averaging for all system configurations s, users u and journeys i. Conditional expectations when fixing a system configuration, or a user, or an instance, or pairs of them, are also possible, depending on what is known about i, s and u.

3.1. Predictability formulation

Predictability of an AI ecosystem is driven by two factors: the intrinsic uncertainty (entropy) of its conditional validity distribution p, and the representational power of the predictors $\hat{p}$ we can bring to bear. To compare the actual validity with the predicted validity we will use strictly proper scoring rules, any function S that minimises its value when the predicted probability distribution $\hat{p}$ is the same as the ground truth p. For instance, the Brier score ($B := \frac{1}{N}\sum_t (f_t - o_t)^2$ for f_t and o_t the forecast and observation respectively) and log-loss ($L := -\frac{1}{N}\sum_t (o_t \log f_t - (1 - o_t)\log(1 - f_t))$) are proper scoring rules. As we are using losses, instead of using *Predictability* ($\mathbb{P}$) as in the previous section, we here define the *unpredictability* $\mathbb{Q}$ of an ecosystem as the best possible forecasting validity error that any predictor in the considered family can achieve. Concretely, let $\mathcal{F}_b = \{\hat{p}\}$ be a family of predictors whose training and evaluation

Table 3
Examples of situations where we need to predict the outcome of an AI ecosystem. For each example, one can create validity predictors that take the input features to anticipate a given output or validity indicator. Many examples are based on existing literature with actual experimental results, while others are formulated to cover different levels of granularity that are yet to be explored.

Examples	Input Features	Validity Indicators
E1. Self-driving car trip: A self-driving car is about to start a trip to a mountain. The weather is rainy and foggy. The navigator is instructed to use an eco route and adapt to traffic conditions but being free to choose routes and driving style. Before starting, the passengers want an estimate on whether the car will reach the destination safely. It is well-known that many factors, such as weather conditions, affect self-driving cars [17,18].	The route, weather, traffic, time, trip settings, car's state, …	Success in safely reaching the destination.
E2. Cost-effective data wrangling automation: A data science team attempts to automate data wrangling, a monotonous and laborious data preparation task that formats data from text, forms, spreadsheets, etc. The team plans to use four LLMs (that differ in cost) to assist them. However, they want to identify the cheapest correct LLM for each individual use case—this kind of dispatching being known as 'routing' [51]. If no LLM is predicted to be correct then the task will be rejected, thereby optimally saving cost and problems. This case is developed in [52].	Meta-features of the textual input instance, architectural information of the LLMs, …	Accuracy of the output for the requested data wrangling task.
E3. Content moderation on a multimodal LLM: An AI provider is releasing a multimodal LLM for public use. To ensure safe deployment, the company implements a content moderation system that inspects every prompt to predict if the LLM will output content that violates safety policies (toxic language, pornographic images, discrimination, unlawful or dangerous material, etc.), and rejects the prompt if it is the case. See examples in [53–55].	Information of the input prompt, safety scores of the LLM's past responses to similar prompts, …	Safety score of the output according to safety policies.
E4. Balanced reliance in human-chatbot interaction: One high-school student is using a new chatbot to help them with their homework but would like to avoid over-reliance or under-reliance. Better human mental models of the chatbot's error boundaries can be developed on the continuous interactions, which may help the student accurately anticipate the chatbot's failures on future homework. Related examples can be found in [11,23,56].	Characterisation of the homework, chatbot's past performance on similar exercises, …	The accuracy of the chatbot on the exercise.
E5. AI agents in an online video game: In a popular online e-sports competition, several AI agents are to be used to form teams. The game developers have previously tested several multi-agent reinforcement learning algorithms. The developers want to anticipate the outcome of the next game against their competitors, based on the characterisation or information of the chosen algorithms and team members. See related examples in [57,58]	Team line-up (own and other teams), match level, …	Match result (score) indicating win, loss or tie.
E6. Training the next frontier LLMs: The pre-training of LLMs is extremely expensive. A technology company aims to predict the performance scores (over a set of AI benchmarks) of a class of hypothetical LLMs via scaling up with an optimal combination of computational resources (training compute, tokens and model size). Examples: [53,55,59].	Amount of training FLOPs, number of tokens, number of model parameters, …	Downstream performance (e.g. accuracy) of the new LLM.
E7. Marketing speech generation: A request is made to a LLM to generate a marketing speech based on an outline. The stakeholders expect the content of the speech to be original, or even surprising. What needs to be predicted is whether the system will generate a speech along the outline, containing no offensive or biased content, and effectively persuading the audience to purchase the product. Models of pitch success have already been explored [60].	Speech outline, audience demographics, potential restrictions, …	Number of people in the audience being persuaded to purchase the product over several months.
E8. Video generation model training and deployment: Drawing from evaluations of prior music video generation models for a social platform, the goal is to predict the quality of several upgraded AI models, as a function of model size, training data, learning epochs, etc., to optimise resources. The extent to which the videos will conform to content moderation standards also matters. This is an example of video generation scaling laws [61], but if done at the instance level and including generation time as input, this can later be used for routing or single-query optimisation.	Quantity of videos, amount of compute, number of epochs, architecture specifications, …	Quality and compliance of generated videos, according to the judgements of recruited experts.
E9. AI assistant in software firm: A software company plans to deploy a new AI assistant to help programmers write, optimise and document their code. The question is how much efficiency (e.g., work hours of saving) the company can get in the following six months [62]. An example is that of [63], identifying what characteristics of a software project are suitable for LLMs, focusing on predictability at the level of task (not instance-level).	Information of tasks, AI assistant details, user profiles,…	Efficiency metric (work hours saved).

Table 3
Continued.

Examples	Input Features	Validity Indicators
E10. LLM user dependency: Users interacting with an LLM for a long period of time adapt their sequence of requests to expectations of previous successes and failures. Scientists aim to *monitor* and anticipate the user's future dependency to the LLM, which is measured by a complex metric that takes into account the loss of independent ability in problem-solving, and other factors. This can be done via long-term studies, proxy signals, intermediate metrics, surveys or simulations (see, e.g., [64–67]).	Sequence of requests from the user, the user's profiles, …	Dependency level (score) judged by an expert-crafted questionnaire.

Table 4
Key distinctions between predictable AI and related areas.

Related Area	Objectives	Differences
Explainable AI	Explainable AI aims to find out what exactly led to particular decisions or actions, and give justifications when things go wrong [2,68–70]	Predictable AI aims to *anticipate* indicators that are observable. Predictable AI is ex-ante, not even requiring to run the AI system to anticipate the outcome. Explainable AI is ex-post; for instance, LLMs can simply mimic human-like explanations by chain of thought or a posteriori.
Interpretable AI	Interpretable AI tries to map inputs and outputs of the system through a mechanistic approach [71–73]	Predictable AI does not aim to build a mechanistic input-output model of the system, but to build a meta-model (predictor) that maps a possibly different set of inputs to specific validity indicators such as performance or safety.
Meta-learning	Meta-learning (i.e., learning to learn) relies on average past performance for future predictions, usually to find the best algorithm or hyperparameters for a new dataset and task [74, 75].	Predictable AI focuses on ways to obtain nuanced predictions that are specific to particular systems but also for each instance and context of use.
Uncertainty estimation	Some AI models output probabilities of success, with calibration and uncertainty estimation techniques focusing on the quality of these probabilities [40,76–79].	Predictable AI is not limited to predicting success, next token or action probability, and the prediction can be done before running the system. Also, unlike uncertainty self-estimation that usually depends on particular machine learning techniques, Predictable AI can be used with any AI paradigm (from deep learning to planning algorithms).
Verification and validation	This process aims to thoroughly verify and validate the system, respectively ensuring it is correct (meets the specification) and ultimately valid (meets the intended purpose) [41, 42].	Predictable AI does not look for full verification or validation of the system, but for probabilistic estimates of those areas where the system meets some indicators such as success or safety.
Causal AI	Causal AI aims to construct causal models with machine learning algorithms, such as causal representation learning [80], and make inference beyond the i.i.d. data assumptions [81].	Predictable AI does not necessarily model the causal mechanisms behind the behaviour of AI, nor does it assume the key variables of the ecosystems are isolated within a causal diagram. Causal models usually target the output but not necessarily the validity.
AI Fairness	AI fairness is about detecting and mitigating discrimination and bias on protected attributes [82], but not on predicting that bias. It focuses on ensuring equal treatment and opportunities across diverse populations.	Predictable AI anticipates validity outcomes, such as bias, either at global level (for various populations) or at granular level (for each instance). Traditionally, bias has been estimated at the populational level, or blocked with moderation filters once given the output, but rarely predicted for individual instances.

cost is bounded by budget b. Unpredictability $\mathbb{Q}$ is the minimal expected S-loss on V that any predictor in $\mathcal{F}_b$ can attain, averaging over draws from p and $\mathcal{E}$.

$$\mathbb{Q}(p, \mathcal{E}, \mathcal{F}_b) := \min_{\hat{p} \in \mathcal{F}_b} \mathop{\mathbb{E}}_{\substack{\langle i,s,u \rangle \sim \mathcal{E} \\ v \sim p(V \mid \langle i,s,u \rangle)}} S(\hat{p}(V \mid \langle i, s, u \rangle), v) \tag{2}$$

For example, imagine $\mathcal{F}_b$ is the family of logistic-regression models that use at most b input features. In our running self-driving car illustration (Fig. 1), journeys A and B are largely determined by a single feature, say "windingness", so even with $b = 1$ the unpredictability $\mathbb{Q}(p, \mathcal{E}, \mathcal{F}_1)$ is low. Journeys C and D also depend on "fogginess" so you need $b \geq 2$ to drive down $\mathbb{Q}$. Yet E and F remain unpredictable under any small linear or logistic model on these features.

Eq. 2 can be read as follows: given a joint distribution over $\langle i, s, u \rangle$, how hard is it to predict the validity of the outcome v in expectation? We can narrow it further if we fix a specific AI system and a specific user. For example, consider the system GPT5 with fixed context (no memory) and a particular user Alice (also with no memory between requests), then $\mathbb{Q}$ would be the level of unpredictability of the validity of the responses that GPT5 provides over the distribution of instances that Alice is requesting. In

Table 5
Examples of situations (described in Table 3) where we need to predict the outcome (validity) of an AI ecosystem. According to the formalisation of unpredictability, the examples are characterised by different levels of granularity on I, S, U, and V (the first three columns correspond to the input features for the alt predictors to produce the validity in the last column).

Examples	I (instances)	S (systems)	U (users)	V (validity)
E1. Self-driving car trip	Single journey	Individual self-driving car	Human passengers	Safe arrival (binary outcome)
E2. Cost-effective data wrangling automation	A data wrangling task	Single language model	Data scientists	Accurate output for the requested data wrangling task
E3. Content moderation on a multimodal LLM	A single prompt	Multimodal LLM	User	Safe output that does not violate safety policy
E4. Balanced reliance in human-chatbot interaction	Query history	Chatbot	Students	Reliable use of the models by the users in the short term
E5. AI agents in an online video game	Online video game	Set of AI agents	N.A.	Game outcome (current or future)
E6. Training the next frontier LLMs	A collection of downstream tasks	A class of hypothetical LLMs	Human users	Accuracy on downstream tasks
E7. Marketing speech generation	Single outline for a speech	Text generator	Group of potential customers	Sales impact (considering reputation, etc.)
E8. Video generation model training	Each prompt to be turned into videos	Video generation model	Social network users	Feedback integration (likes, rewards) and future video outcomes
E9. AI assistant in software firm	Each programming task	AI assistant	Programmers	Efficiency (work hours saved) and code quality (robustness, bugs)
E10. LLM user dependency	Each request	Language model	User	Dependency metrics (loss of independent ability, mental health impacts)

general, varying levels of granularity are possible. For example, one may set a specific AI system (GPT5) interacting with a large set of users, where $\mathbb{Q}$ would then correspond with the level of unpredictability of the validity of GPT5's responses over the distributions of instances that all users are requesting. For such setups, one may build predictors fed with the features of individual instances, systems or users $\langle i, s, u\rangle$, depending on the level of granularity for a specific application interest.

Let us explore again the examples in Table 3 but formally characterising the notions of I (instance), S (system), U (user) and V (validity). We provide this characterisation in Table 5. Here, entries vary in complexity, with some capturing complex ecosystems, while others single interactions between an AI system and a user for a single instance. For example, with the case of *"E1. Self-driving car trip"*, we do not consider the full ecosystems of a self-driving vehicle fleet, but rather just individual journeys of a single car, independent of previous trips. However, each of these examples can be viewed through different lenses of time horizon h. A car journey in E1 can be viewed as an atomic event, with a single outcome, or as a sequence of events that combine to form the entire journey. In the former case, Eq. 2 is sufficient[2], in the latter, the more general formalisation from Appendix A is required.

We can now revisit the difference between what we refer to as the 'base problem' and the 'alt problem'. The base problem is the original task the AI system is designed to solve, (e.g., finding a set of actions that safely drives the car to the destination). The system s maps instance i to output or behaviour $s(i)$. The alt problem, on the other hand, attempts to build a meta-model that can estimate whether s will succeed at any given instance, i.e., it maps system s, instance i and user u to a validity estimate. To do this, one can estimate a distribution $\hat{p}(V \mid \langle i, s, u\rangle)$ (e.g., through machine learning, for which we will need validity annotations on a dataset to train the alt predictor). Importantly, the alt problem does not predict the output or behaviour of the system, but instead predicts its validity. This differs significantly from other externalised meta-frameworks such as Guaranteed Safe AI [83], which models the mapping between inputs to outputs (the base model), the mapping between outputs and outcome (a world model) and a mapping between the state and the reward (a reward model). This requires solving the first two—and hard—columns in Table 1. The alt problem is more anticipatory and simpler, mapping inputs to validity directly. The alt problem frames assurance in a space that is no longer cornered between designing systems that are *provably correct*, as in other areas of computer science, or that are *probably correct* (or probably approximately correct [84]), as in the traditional aggregative evaluation in machine learning. Instead, we extend the outcome indicator beyond correctness (safety, fairness, user satisfaction, etc.) and look for its anticipation, shifting the paradigm to *Predictably Valid AI*.

In general, the distinctive trait for considering an AI ecosystem "predictable" is the possibility of having a reliable method that predicts validity from key indicators, by minimising the S loss. This raises the question of what considerations are needed when framing the alt problem such as what to predict, how to predict, and who does the prediction, topics that we address in the following subsection.

[2] As $\hat{p}$ predicts at instance level, it can be used to derive aggregate predictions. Similarly, worst-case or best-case situations (journeys) can be found by applying $\hat{p}$ to a set or distribution of cases, or calculated if $\hat{p}$ is invertible, analytically or by optimisation.

3.2. Framing predictability

Predictable AI aims at any validity indicator that can be reliably anticipated and can be used to determine when, how or whether the system is worth being deployed in a given context. Clear examples of these indicators are *correctness* and *safety*, as measured by certain metrics; but virtually any other indicators of interest, such as *fairness*, *rewards*, *game scores*, *energy consumption* and *response time* could be subject to prediction.

This notion of indicators is similar to that of "property-based testing" in software engineering [85] and recently adapted to AI [86]. However, the focus of Predictable AI is to anticipate the values of these indicators (under what circumstances the system is correct, safe, efficient, etc.) rather than to test or certify that they always have the right value (always correct, safe, efficient, etc., under all circumstances). In other words, predictability can make a non-robust system useful, if we can anticipate its *validity envelope*, the conditions under which operation is predicted to be valid. A formal definition of these envelopes is given in Appendix A.

Apart from determining what is to be predicted, we must also characterise how the alt problem should be framed depending on several aspects, which we call the *predictability framework*:

- **Input Features**: These are denoted by the definition of each of i, s and u as parametrised vectors, with combinations of input features only observed with some combinations of system features. This offers sweet spots beyond the limited amount of input features that the base system usually works with. For instance, a predictor modelling the outcome of the base system can take advantage of additional information of the task i (e.g., meta-features like instance complexity, presence of noise, embeddings, etc.). It can also take the characteristics of other AI systems (e.g., if other more powerful systems fail on the same or similar tasks, this base system may fail too).
- **Anticipativeness**: The predictors can either be anticipative or reactive. Anticipative predictors are run before the system is used, (i.e., the output o is not used to predict the validity indicator). This is necessary, e.g., when determining whether an agent will perform undesirable actions to fulfil a command before it starts operation. In contrast, for certain contexts, we may also consider reactive predictors (validators) that predict the indicators after the system has been run but not yet deployed, adding o to Eq. 2, i.e., $\hat{p}(V \mid \langle i, o, s, u\rangle)$. Examples of validators include content filters or verifiers [87]. Deciding after having seen the output is easier, especially for safety indicators, but could be unsuitable depending on the kind of system, costs, safety or privacy.
- **Granularity**: This determines whether the validity predictions are performed for individual instances, systems and users, or aggregated in certain ways. For instance, predictions can be made at the 'instance level', for the validity of a single input or event $\langle i, s, u\rangle$, or at the 'batch level', as an aggregate for a set of inputs (benchmark metrics are a good example of this, but modelling rare events requires more than extrapolating an average [88]). Similarly, we can make predictions for a specific system or user, or larger-scale predictions as an aggregation of multiple systems or users. The same predictor can navigate different granularities using aggregation and disaggregation techniques.
- **Prediction horizon**: If the actions of an AI system are not considered atomic (explored in Appendix A), then the prediction horizon h could vary. This enables tasks to be broken down into multi-step decision-making processes, for more granular predictability. Both short and long time horizons can draw on recent data inputs or on historical data and trends. The time scale, in conjunction with the granularity, may be segmented and aggregated into finer or coarser periods. Forecasting the future progress in AI systems (e.g., through scaling laws), the technology (e.g., through expert questionnaires) or their impact (e.g., on the work market) is variously difficult, but trends for longer horizons are seen at larger scales, such as predicting the use of compute or energy of AI technology as a whole [89–92].
- **Hypotheticality**: This is represented by the possibility of interrogating $\hat{p}$ such that it can extrapolate about hypothetical systems that do not exist or have not been seen. Interrogating these models is especially useful before building a system (e.g., *"E6. Training the next frontier LLM"* in Table 3) or when deciding some hyperparametrisations or options for deployment (e.g., *"E8. Video generation model training and deployment"*). This also allows us to determine if an AI ecosystem has solvable or safe solutions within the parameters of some current AI technology.

Finally, we identify three distinct ways of predicting the validity of AI ecosystems, by considering who makes the prediction: humans, the base systems themselves, or an external predictive model, prompted or trained using empirical evaluation data. These three options can be used at any level of granularity and time scale. We now discuss concrete examples of each of these three options.

First, human predictions about an AI system's validity indicators can be useful at the instance level. This is usually referred to as human oversight or human-in-the-loop [93]. Such predictions can be anticipative (e.g., users often refrain from certain queries or commands fearing poor results) or reactive (e.g., users can filter out some outputs after the system has been run). The importance of humans predicting AI (and how good humans actually are at it) has been studied recently, especially in the context of human-AI performance [56,94], human-like AI [32,95–97], people's ability to predict a chatbot's errors [23], and the concordance between human expectation and language model's errors [11]. Human predictions about AI ecosystems have been elicited using expert questionnaires [45,98,99], extrapolation analyses [100], crowd-sourcing [101] or meta-forecasting [102]. Another, as-yet underexplored possibility would be to harness the benefits of prediction markets [103] and structured expert elicitation methods [104].

Second, many machine learning systems come with self-confidence or uncertainty estimations [40,78]. These estimations can be interpreted as the system in question predicting its own likelihood of success. If well-calibrated, these estimates can be powerful predictors of performance; [52] make use of the self-confidence of four variants of GPT-3 to assess how good these LLMs are in self-estimating their own success. However, the models may not be well calibrated. For example, LLMs were becoming better calibrated [105,106], but subsequent fine-tuning and reinforcement learning from human preferences were shown to significantly degrade this calibration [53]. Even in cases where calibration is good on the target distribution, there are limitations to predicting in

Table 6
Predicting the success of agents in the Animal AI platform using five different approaches [21]. From left to right: (i) the majority class prediction, (ii) global accuracy extrapolation, (iii) each agent's accuracy extrapolation, (iv) a predictive model, C5.0, using all instance features and agent id as inputs, and (v) same as iv but only using the three most relevant instance features (reward size, distance, and y-position) and the agent id.

	Maj. (1)	G.Acc.	T.Acc.	All + A	Rel + A
Brier score ↓	0.453	0.248	0.176	0.148	0.154

this manner. This approach is limited to what the system has seen (i.e., a system only has access to its own training data, not those of other systems, which may provide with additional information). In addition, uncertainty estimators are built at the level of token or specific action when using general-purpose AI systems, incapable of reliably anticipating task success before starting it. Further, there are cost implications, as the base system must be run for each instance to obtain the self-estimation, such as log probabilities per token in language models.[3] Furthermore, leaving the system to predict its own performance creates a conflict of optimisation goals, potentially trading performance with uncertainty estimation quality. There may even be a direct feedback loop between the model and the user, which has been identified as one of the main drivers of misaligned behaviour, such as deception and manipulation of humans [34,107,108]. Hence, while self-estimation can be an option [109], it is generally less versatile than building independent predictors $\mathcal{F}_b$ with separate entities like humans or external predictive models. Also, we can build as many alt predictors for a battery of validity indicators, whereas self-confidence is generally restricted to performance.

Third, the final option is to train a predictive model from observed data about the validity of the base model. A straightforward way of doing this is by collecting test data about systems and task instances (and possibly users) and training an "assessor" model [52, 110–112] or a moderation filter [53–55], a predictive model that maps the features of inputs and/or systems to a given outcome (e.g., validity or safety). An alternative way is to identify the demands or difficulty measures of the task and build a model that relates demands and capabilities to performance, using domain expertise [113–116]. This approach is often called capability-oriented or feature-oriented evaluation [117,118] and has the potential to be reverse-engineered to explain instance-level hardness [119]. Probes can also be used for predicting success or abstention [120]. All these models can be used to predict how well a system is going to perform for a new task instance based on task demands and system capabilities. In both cases (assessors and capability-oriented evaluation), instance-level experimental data is needed [121]. Human feedback is another important source of data, often used to build reactive predictors through reinforcement learning (RLHF) or other techniques [122–126]. Predictive models can also be built at higher levels, with distributional or aggregated data [88,127–129]. For instance, the use of scaling laws to anticipate model performance on benchmarks [53,55,59,130–132] is a very popular contemporary approach. Still, other predictive models can be built from aggregate indicators [133–135] at high levels of abstraction, as is common in the social sciences and economics. Finally, this external predictor does not need to be necessarily trained; language models have been used to predict validity indicators of other models without training or fine-tuning, just by prompting [136].

4. Scenarios

To shed more light on the above aspects framing predictability, we explore seven realistic scenarios that vary in scope and focus, ranging from predicting performance of base systems on specific tasks to analysing the broader "scaling laws" in neural models. We will find the three types of predictors (humans, self-estimation from the base systems, and external predictive models) in these examples.

In the **first** scenario, the objective is to predict the performance of an AI agent in a new navigation task, using information about the behaviour of the agent itself, other agents approaching similar tasks and the characteristics of the tasks. In particular, [21] consider navigation tasks in the 'AnimalAI Olympics' competition [137,138], using the results of all the participants. Their goal is to anticipate success (1) or fail (0) for each task. To that purpose they use five distinct approaches ranging from predicting the most frequent class to building a predictive model $\hat{p}(V \mid i, s)$ using the most relevant instance i and system s features. As we can see in Table 6, the last approach (Rel + A), using the three most relevant instance features (reward size, distance and y-position) together with a system feature (agent ID), can predict task completion with a Brier score of around 0.15, demonstrating that a choice of a small set of relevant features can lead to an effective predictor.

The **second** scenario comes from example *"E2. Cost-effective data wrangling automation"* in Table 3. Zhou et al. [52] focus on the task of automating data wrangling using the results from four variants of GPT-3 models under distinct few-shot setups. They attempt to anticipate and reject instances for which GPT-3 models will predictably fail, to avoid unnecessary costs. To this end, they build a small assessor model (using a random forest approach) as the predictor $\hat{p}$, fed by the details of the instances i (e.g., meta-features, number of shots, etc.) and the base systems (e.g., model size, architecture, etc.), that can make reliable predictions of the performance of the base systems s (GPT-3 models). The target is thus to predict $V = 1$ (success) or 0 (failure). They also compare the predictive power of $\hat{p}$ with a baseline formed by the self-estimation of base systems. The results are shown in Table 7, where they find good

[3] Such cost implications also occur with reactive predictors due to the requirement of obtaining the output from the base system

Table 7
Comparison between Brier score (BS) of the assessor's predictive power and the self-estimation baseline from GPT-3 models [52]. The average accuracies (with standard deviation) across different numbers of shots from base models, GPT-3 variants, on the data-wrangling task are also presented.

Base model	Base model's Acc. ↑	BS of self-estimation ↓	BS of $\hat{p}$ ↓
GPT-3 Ada 350M	0.524±0.232	0.122	0.144
GPT-3 Babbage 1.3B	0.580±0.240	0.116	0.141
GPT-3 Curie 6.7B	0.625±0.244	0.108	0.130
GPT-3 Davinci 175B	0.689±0.253	0.096	0.125

prediction quality (as measured by Brier score) from both $\hat{p}$ and self-estimation in predicting performance of all base models. While self-estimation is slightly better, the external alt predictor $\hat{p}$ does not need to run the LLM when the rejection rule is enabled, saving computational cost. By rejecting those instances that were predicted with an estimated probability of success lower than 1%, 46% of the failures were avoided, at the cost of only rejecting 1.5% of correct answers. They also report that various meta-features of the task instances and architectural details of the base systems can augment the predictive power of $\hat{p}$, highlighting the relevance of including features beyond what the original task (base problem) considers [52, Table 3]. Today, runtime parameters such as reasoning time could also be used.

A **third** scenario explores how easy it is to find features for $\langle i, s, u\rangle$ that are predictive, and how good humans are at finding them. Here, humans act as predictors $\hat{p}(V \mid \hbar(i), s)$ where the input feature is the user's perceived difficulty $\hbar$ for instance i. V is the actual LLM correctness for (i, s). For each instance, we can view the problem as estimating $\mathbb{Q}(p, \mathcal{E}, \mathcal{F}_{\text{human}})$. Zhou et al. [11] found that human-estimated difficulty is a good predictor of performance in LLMs (Fig. 2). This indicates that future LLMs could use human difficulty to determine when to abstain from providing an answer [139]. Furthermore, humans can also use it to reject the model's output for difficult questions, acting as a predictor. While this is promising for both machine and human oversight, Zhou et al. [11] noted that in practice humans do not leverage this difficulty well when spotting and rejecting possible errors, corroborating previous observations about humans being unable to determine where LLMs fail [23]. However, the predictability is there, ready to be exploited.

Relatedly, in a **fourth** scenario, Dreyfuss and Raux [140] investigate whether ordinary people can tell in advance when an AI will succeed or fail, and how those beliefs shape delegation decisions. Here the predictive function is again $\hat{p}(V \mid \hbar(i), s)$ (for various i), with V as before. They run two controlled online experiments plus a real-world test with a parenting chatbot. Across all settings, participants judge the AI through a human lens: a single slip on a question they consider "easy" makes them distrust the system, while a single triumph on a "hard" one makes them over-trust it-even though those cues say little about the model's true strengths and weaknesses. Because of this human-projection bias, people's probability forecasts are poorly calibrated (average Brier score ≈ 0.25 versus 0.15 for a calibrated predictor, where the observed BS is an empirical proxy for $\mathbb{Q}$ for the family of predictors generated by the human participants) and their adoption choices become extreme: they either hand every task to the AI or refuse to delegate anything, ignoring the genuinely optimal mix. The study reinforces the observation from [11]: lay users do not seem good at leveraging difficulty to predict LLM performance. They also show that human-like interface may amplify errors of lay users, and argue for adding explicit confidence signals or external assessor models to keep users' intuitions in check. (Table 8, Table 9)

Our **fifth** scenario is ADeLe (AI Evaluation with Demand Levels), a fully-automated framework by [115] that enables the construction of robust predictors for LLM performance at the instance level. Concretely, the authors build an assessor model (using random forests as predictor family) fed with the demand levels of 18 abilities scales (e.g., quantitative reasoning, knowledge, comprehension, etc.) of individual task instances. These features allow them to *predict* the performance of 15 LLMs from OpenAI, Meta and DeepSeek that varied in model size, training methods, embedded optimised chain-of-thought reasoning, etc. The assessor is trained on past performance records of LLMs and is able to robustly predict LLM performance on new instances in seen (in-distribution) and unseen (out-of-distribution) tasks with nearly perfect calibration and high discrimination power. This high predictability can be used for better routing methods to choose what model to use [51], determining safety operating areas where assurance is guaranteed and anticipatory reject rules when harm or cost is anticipated [52,141].

Our **sixth** scenario focuses on autonomous agents performing complex, long-horizon software engineering and cybersecurity tasks, as analysed in the HCAST (Human-Calibrated Autonomy Software Tasks) framework [91,92]. In this setting, the key feature for anticipating success is "human completion time": the time a domain expert would need to solve the same task instance. The authors show that an agent's validity (i.e., its probability of autonomous success) is well captured by a logistic function of the *log* of its duration (Fig. 3). This high level of predictability derives from a geometric series on the probability of failure at each step, assuming steps have similar degrees of (in)dependence. Conversely, the logistic function can be seen as a cumulative probability of failure over time; for example, we can determine whether a model is likely to fail before a specific temporal threshold, such as four minutes. Building on this, Appendix A extends the formulation of predictability to account for a time horizon, as in this scenario.

The **seventh**, final scenario focuses on the so-called "scaling laws" [59], which represent a power-law relationship between the overall performance of language models for a set of tasks and the increase in factors such as model size, dataset size and computational power (see Fig. 4). Here, the input variables are compute, data size and number of parameters. These are proven to be highly predictive for neural models' test loss, with loss linearly decreasing with these parameters (log scale) [59,130]. Thus, we are interested in predicting a validity indicator V (test loss, or accuracy) for a hypothetical system s (characterised by compute, data, and parameters) and possibly instance i (per-task or per-instance prediction). The predictive model is $\hat{p}(V \mid s)$ at the aggregate-level or $\hat{p}(V \mid i, s)$ at the

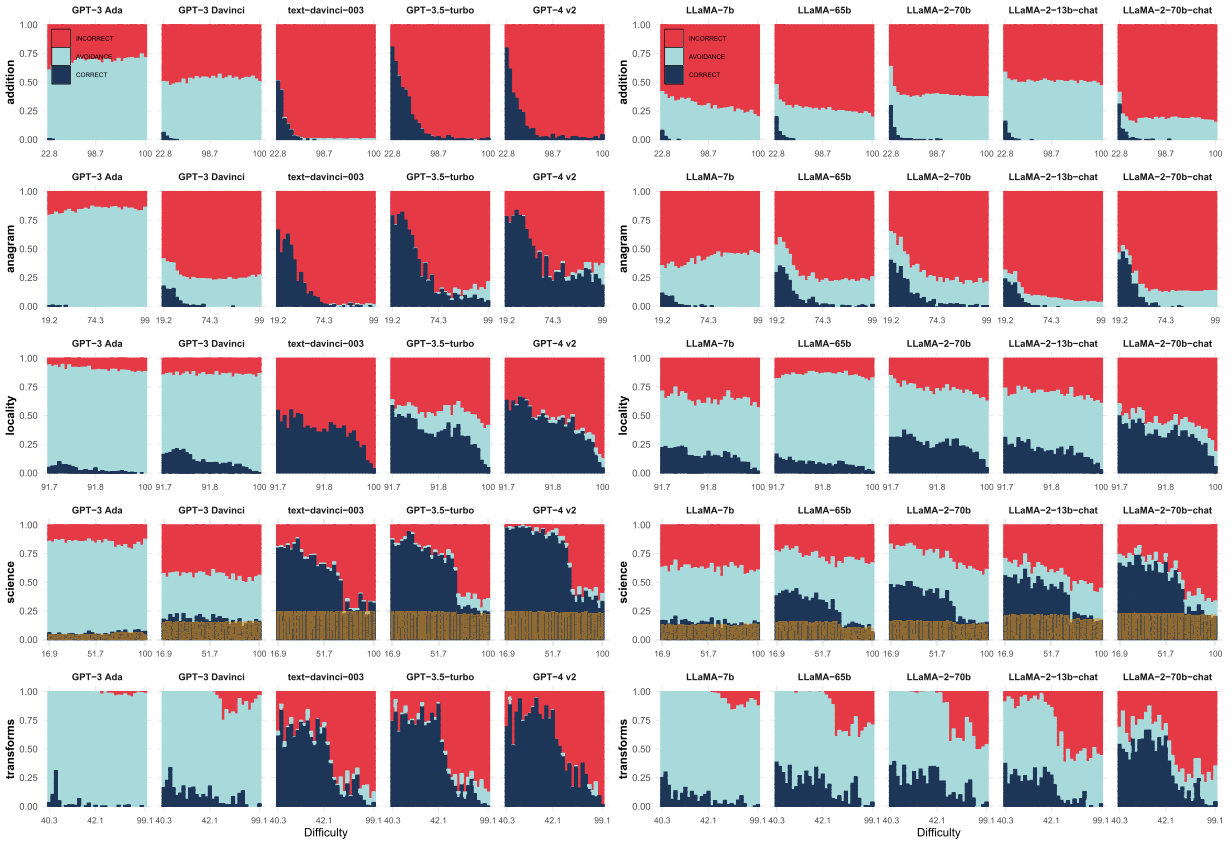


Fig. 2. Performance of a selection of GPT and LLaMA models over human difficulty on the ReliabilityBench benchmark [11]. The values are split by correct, avoidant and incorrect results. The x-axis is split into 30 equal-sized bins, whose ranges must be taken as indicative of different distributions of perceived human difficulty across benchmarks.

Table 8
In-distribution predictability results of 15 LLMs over 20 benchmarks in the ADeLe v1.0 battery from [115]. The first two columns show names of subject LLMs and the overall accuracy of subject LLMs on the ADeLe battery. The remaining three pairs of columns show the AUROC and ECE of three different assessors (RF using demands, RF using average GloVe embeddings, and finetuning LLaMA-3.1-8B).

Subject LLM	LLM Accuracy↑	Demands (RF)		Embeddings (RF)		Finetuning (LLAMA)	
		AUROC↑	ECE↓	AUROC↑	ECE↓	AUROC↑	ECE↓
Babbage-002	0.102	0.786	**0.004**	0.784	0.012	**0.794**	0.026
Davinci-002	0.157	0.774	**0.005**	0.770	0.014	**0.789**	0.032
GPT-3.5-Turbo	0.414	0.811	**0.007**	0.780	0.029	**0.817**	0.052
GPT-4o	0.713	**0.882**	**0.014**	0.852	0.041	0.879	0.039
OpenAI o1-mini	0.770	0.860	**0.011**	0.821	0.023	**0.861**	0.041
OpenAI o1	0.843	**0.853**	**0.011**	0.810	0.025	0.848	0.031
LLaMA-3.2-1B-Instruct	0.216	0.785	**0.006**	0.759	0.014	**0.788**	0.041
LLaMA-3.2-3B-Instruct	0.378	0.813	**0.008**	0.782	0.028	**0.822**	0.048
LLaMA-3.2-11B-Instruct	0.463	0.820	**0.009**	0.793	0.034	**0.828**	0.055
LLaMA-3.2-90B-Instruct	0.645	**0.860**	**0.012**	0.832	0.042	**0.860**	0.042
LLaMA-3.1-405B-Instruct	0.683	**0.870**	**0.011**	0.831	0.040	0.864	0.040
DK-R1-Dist-Qwen-1.5B	0.353	0.781	**0.014**	0.749	0.028	**0.797**	0.052
DK-R1-Dist-Qwen-7B	0.555	0.813	**0.015**	0.788	0.039	**0.821**	0.051
DK-R1-Dist-Qwen-14B	0.698	0.828	**0.013**	0.796	0.031	**0.829**	0.044
DK-R1-Dist-Qwen-32B	0.748	**0.841**	**0.013**	0.799	0.031	0.839	0.045
Weighted Average	—	0.839	**0.011**	0.805	0.032	**0.840**	0.043

Table 9
Task out-of-distribution predictability results from [115] (all other things equal to).

Subject LLM	LLM Accuracy↑	Demands (RF)		Embeddings (RF)		Finetuning (LLAMA)	
		AUROC↑	ECE↓	AUROC↑	ECE↓	AUROC↑	ECE↓
Babbage-002	0.102	**0.751**	**0.007**	0.727	0.019	0.719	0.046
Davinci-002	0.157	0.741	**0.007**	0.703	0.025	**0.746**	0.055
GPT-3.5-Turbo	0.414	**0.795**	**0.020**	0.719	0.032	0.773	0.088
GPT-4o	0.713	**0.852**	**0.023**	0.789	0.073	0.831	0.067
OpenAI o1-mini	0.770	**0.837**	**0.021**	0.751	0.038	0.814	0.068
OpenAI o1	0.843	**0.811**	0.033	0.730	**0.030**	0.761	0.101
LLaMA-3.2-1B-Instruct	0.216	**0.733**	**0.026**	0.671	0.033	0.732	0.081
LLaMA-3.2-3B-Instruct	0.378	**0.791**	**0.016**	0.724	0.020	0.780	0.084
LLaMA-3.2-11B-Instruct	0.463	**0.799**	**0.022**	0.733	0.037	0.783	0.106
LLaMA-3.2-90B-Instruct	0.645	**0.834**	**0.021**	0.763	0.068	0.809	0.050
LLaMA-3.1-405B-Instruct	0.683	**0.843**	**0.023**	0.766	0.067	0.811	0.060
DK-R1-Dist-Qwen-1.5B	0.353	0.757	**0.019**	0.700	0.029	**0.764**	0.071
DK-R1-Dist-Qwen-7B	0.555	**0.790**	**0.018**	0.735	0.042	0.776	0.083
DK-R1-Dist-Qwen-14B	0.698	**0.808**	**0.018**	0.737	0.054	0.772	0.085
DK-R1-Dist-Qwen-32B	0.748	**0.812**	**0.026**	0.739	0.057	0.793	0.063
Weighted Average	—	**0.810**	**0.022**	0.740	0.047	0.788	0.075

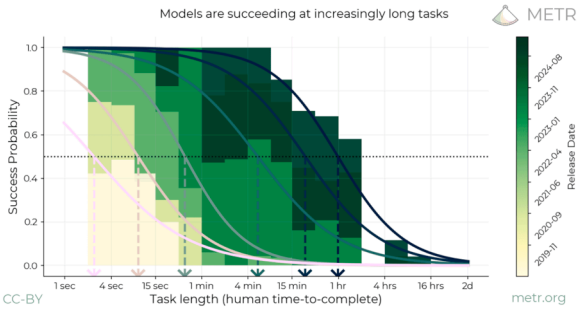


Fig. 3. Logistic curves derived in [91,92] for success probabilities (validity scores) on software tasks that would take humans a certain amount of time to complete. .

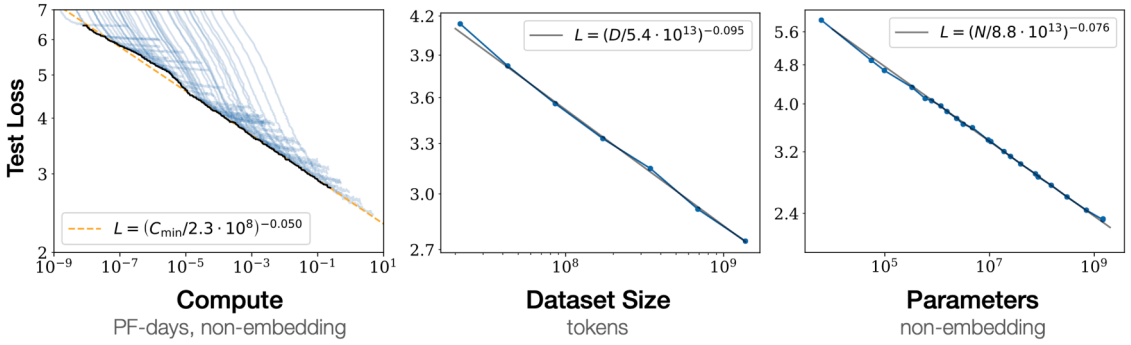


Fig. 4. Scaling laws of neural models [59]. The test loss is predictable from the compute used during training, the training dataset size and the number of parameters of the model.

instance-level. This scenario is a clear case of long-horizon hypotheticality (i.e., evaluating $\mathbb{Q}$ for unseen s) that is usually addressed with coarse granularity (how a hypothetical model will perform on a dataset [129]), but there is increasing interest in building predictors at the instance level and derive scaling laws from them [115,142]. This anticipation at the instance level is even more relevant as 'reasoning' models can be parametrised by the 'thinking budget', as they "improve with more thinking time" [143]. If we can anticipate that the model can give us a good result by thinking during 5 seconds, why should we give it 100 seconds? Conversely, if we anticipate it cannot give us a good answer, why should we spend all these costly thinking seconds?

These scenarios emphasise the relevance of the input features and also share an anticipative character. They differ on the prediction horizon, and are situated at different levels of aggregation: the local, fine-grained prediction at the instance level (for the first five scenarios); the cumulative element of time in the sixth scenario, and the global, coarse prediction for massive benchmarks

(the seventh scenario). There are many intermediate areas where predictability has not been explored. All these scenarios also highlight the difference between predicting performance of a specific AI system and making a more general prediction about a class of hypothetical (not yet trained) AI systems, as represented by the scaling laws and other common uses of expressions such as 'predicting AI capabilities' [144–146]. This exploration of intermediate levels, varying scales and different types of validity indicators is fundamental to understanding possibly confounding effects of the aggregation, such as a biased selection of the relevant input or output variables until predictability is found [147].

5. The trade-offs

We advocate for a paradigm shift where the design, selection and use of predictable AI systems are prioritised. However, there is a tension between predictability and the quality of the base systems, because a model that always fails is fully predictable. There is further tension in the effort that must be expended to minimise Eq. 2. Ultimately, what we would like is useful AI systems and a good predictor $\hat{p}$ for the resulting validity indicators of the AI ecosystem.

The first idea for building more predictable AI systems, especially machine learning models, may be based on keeping them simple (e.g., a set of causal rules instead of a complex black box model). However, this only entails behavioural predictability but may not ensure more validity predictability. For instance, as we saw in Table 2, a classifier s that always predicts the same label is very simple and very explainable, but predicting where it fails, the $\hat{p}$ problem, would require learning the original classification problem. In general, if the AI models have not captured the epistemic uncertainty of the base problem, this will make the alt problem harder, as this epistemic uncertainty would need to be instead captured by $\hat{p}$.

Making explicit the balance between minimising $\mathbb{Q}$ and maximising expected validity $\mathbb{V}$ is explored in Eq. A.3 in the Appendix. Here we point out some other ways to interpret this trade-off:

- Explore the Pareto between the expected validity $\mathbb{V}$ and reducing the loss S of $\hat{p}$ w.r.t. p. For instance, in the scenario of the Animal AI Olympics seen above, there were some participants, such as 'Sparklemotion', that showed higher accuracy than other weaker participants, such as 'Juohmaru', but much worse predictability [21]. A Pareto plot (x-axis and y-axis equal to accuracy and predictability, respectively) could place "Juohmaru" as preferable.
- In the case of binary validity distribution p, if the predictor $\hat{p}$ is well calibrated, we could set a threshold to determine the proportion of operating conditions $\mathcal{E}_{\hat{p},\tau} = \{e \in \mathcal{E} : \hat{p}(V = 1 \mid e) \geq \tau\}$ that are not rejected and the percentage of these regions that are actually above the threshold of quality, i.e., $\mathbb{E}_{e \in \mathcal{E}_{\hat{p},\tau}} p(V = 1 \mid e) \geq \tau$. In the second scenario, *"E7. Cost-effective data wrangling automation"*, the proportion of operating conditions increased from 55.2% (without rejection) to 69.2% at the cost of rejecting 1.5% of correct answers, with $\tau = 0.01$ [52, Table 4]. If the predictor is well calibrated, this proportion can be further increased by increasing τ, but this will also further reduce the number of correct answers the user receives.
- Instead of setting a threshold, which is very application or context-dependent, we can optimise for some metrics that combine high validity and low rejection using $\hat{p}$, such as the area under the accuracy-rejection curve [148] or extensions beyond classification.

PredictaBoard [149] is a benchmarking framework dedicated to evaluating the predictability of LLMs via score assessors that considers different metrics to navigate the trade-off between the quality of the base model and the assessor. PredictaBoard systematically quantifies how reliably a given assessor can anticipate LLM failures or successes on individual prompts, introducing the notion of a "safe operating zone" by reporting the rejection rate across various error tolerances. This enables a direct and fair comparison among (LLM, assessor) pairs.

Specific solutions can be tailored for each AI ecosystem, but the choice of the validity metric, its maximisation (through better AI systems) and the optimisation of its predictability (through better predictors $\hat{p}$) will be central to the essential challenges and opportunities of Predictable AI.

6. Challenges and opportunities

Characterising the field of Predictable AI allows us to better delineate its challenges and turn them into focal research opportunities rather than scattered efforts. The following list is not exhaustive, but builds on the elements identified in previous sections:

- **Metrics**: Can we use the traditional evaluation metrics for performance, usefulness, safety, etc., or do we need new metrics such as alignment, honesty, harmlessness, helpfulness [150]? What properties make a metric more easily predictable? How do we identify when a system is predictable enough [115,151]?
- **Evaluation data**: What data to collect for training assessors or evaluate their predictiveness [121]? How can we combine human feedback, predictions from different actors [23], results from other systems [152], incident databases [153], meta-feature construction and annotations [154]?
- **Aggregation and disaggregation**: Can different predictability problems at several granularities be bridged, from local, instance-level predictions to global, benchmark-level predictions and vice versa? Is quantification [155,156] the right tool for this?
- **Effective monitoring**: How can we integrate different predictors to monitor AI ecosystems and federate them [157] in case of multiple users and stakeholders? What are the liability implications and how should this be regulated? [158]
- **Reuse of knowledge**: How can we reuse domain knowledge from cognitive science about how humans and animals solve tasks [95, 96,115,137,138] or from what explainable and interpretable AI finds about an AI system?

Regarding all these challenges, and especially the reuse of knowledge, we see opportunities in exploiting the synergies of comparing validity predictability in AI with predictability in other sciences [159–168].

A crucial research niche that is intertwined with the previous list is the *identification of pathways toward improving the predictability* of AI ecosystems. There are several promising methods for this[11]. propose two ways to increase error predictability of LLMs from a human perspective: (1) modify loss functions to penalise errors on easy tasks more heavily than difficult ones so as to enhance the concordance between user difficulty expectations and model errors; (2) use human expectation to make LLMs more epistemically human-like such as abstaining from answering on tasks beyond their capabilities. In a different approach, [169] demonstrate that neural networks self-regularise when given the auxiliary task of predicting their own internal states, making networks more parameter-efficient and reducing complexity, which may increase their predictability. Further, research on mechanistic interpretability has shown promise in learning monosemantic representations from transformer network layers [170]. Researchers can intervene in models to change their internal representations, which could be used to optimise for predictability with respect to a given pair of base model and predictor. On top of these options, other techniques that improve uncertainty estimation [171] could also help external assessors.

Depending on the domain, there are open methodological questions such as who should make the predictions (human experts, the AI systems themselves or an external predictive model), how their predictions should be elicited, and whether Predictable AI can be applied recursively[4] There are further theoretical questions about *how much* can be predicted subject to aleatoric and epistemic uncertainty, and the causal loops involving predictions. Ethical issues, such as privacy of behaviour and responsibility when predictions fail, will require a reunderstanding at both the base and alt levels. In general, many of the above challenges will lead to cross-disciplinary research opportunities.

7. Conclusion, impact and vision

We identified AI predictability as a fundamental, yet underexplored component of many AI desiderata, such as trust, liability, control, alignment and safety. We have shown that predictability is highly intertwined with, but separate from, other important areas such as explainable AI, interpretable AI, meta-learning, uncertainty estimation, etc. Situating it as a field of scientific enquiry in itself recognises common patterns, problems and solutions that have been disconnected in the past, but are now seen under the same umbrella, creating new synergies and opportunities.

A collective shift in focus towards Predictable AI would constitute a profound paradigm shift yielding greater assurances about system performance, safety and deployment suitability. There are reasons to be optimistic about predictability within AI: first, in many other sciences, predictability is a fundamental aspect of operation, and many ideas can be reused; second, there has been enormous progress in predictive techniques, and we expect powerful AI models to be used as predictors of validity of other AI systems in many domains, as we have seen with LLMs; third, there are already promising results by the incipient predictable AI community analysing the predictability of certain AI systems (e.g., LLMs) on certain validity indicators (e.g., performance), as well as improved assessors and other techniques that better predict validity.

One of the key elements in the notion of predictability relies on the granularity of predictions (emphasising instance level anticipation), and the clear distinction of levels ('base' and 'alt'). Both these two levels can be based on AI techniques, but may have very different model families, e.g., some standard LLMs as base models handling complaints, and a family of assessors that can be built from the input feature representation and the validity indicator as the output. The big challenges for the years to come will appear when making AI systems more predictable in the first place and evaluating whether the deployment of the AI ecosystem keeps this predictability at larger scales.

We anticipate that through the framework presented in this paper, more concrete progress can now be made. In particular, the use of machine learning to exploit the increasingly large amounts of evaluation data (benchmark results and human feedback) generated by AI systems holds promise for the development of this nascent field, leading to a landscape of predictably valid AI systems.

CRediT authorship contribution statement

Lexin Zhou: Writing – review & editing, Writing – original draft, Conceptualization; **Pablo A. M. Casares:** Writing – review & editing, Writing – original draft, Conceptualization; **Fernando Martínez-Plumed:** Writing – review & editing, Writing – original draft, Conceptualization; **John Burden:** Writing – review & editing, Writing – original draft, Conceptualization; **Ryan Burnell:** Writing – review & editing, Writing – original draft; **Lucy Cheke:** Writing – review & editing, Writing – original draft, Conceptualization; **Cèsar Ferri:** Writing – review & editing, Writing – original draft; **Alexandru Marcoci:** Writing – review & editing, Writing – original draft; **Behzad Mehrbakhsh:** Writing – review & editing, Writing – original draft; **Yael Moros-Daval:** Writing – review & editing, Writing – original draft; **Seán Ó. hÉigeartaigh:** Writing – review & editing, Writing – original draft; **Danaja Rutar:** Writing – review & editing, Writing – original draft; **Wout Schellaert:** Writing – review & editing, Writing – original draft, Conceptualization; **Konstantinos Voudouris:** Writing – review & editing, Writing – original draft; **José Hernández-Orallo:** Writing – review & editing, Writing – original draft, Supervision, Conceptualization.

[4] This can be done by building a next-level predictor that anticipates how good a given predictor (e.g., assessor) will be at the instance level or globally: this would create a third level, or more if done recursively. One interesting technical question is how much epistemic uncertainty that is not captured at the metalevel can be further captured at the meta-metalevel. This will depend on the power of the family of predictors at each level.

Data availability

No data was used for the research described in the article.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Jose Hernandez-Orallo reports financial support was provided by Future of Life Institute.

Acknowledgements

This work has been supported by Open Philanthropy Long-term Future Fund, CIPROM/2022/6 (FASSLOW), funded by Generalitat Valenciana, and Spanish grant PID2024-162030OB-100 (ROBIN), funded by MCIN/AEI/ 10.13039/501100011033 and ERDF A way of making Europe, Cátedra ENIA-UPV in Sustainable AI Development, TSI-100930-2023-9, and INCIBE's Chair funded by the EU-NextGenerationEU through the Spanish government's Plan de Recuperación, Transformación y Resiliencia, and EUR2024-153548 (PREDAIT) "Towards Predictable AI" from "Spanish Europe Excelencia" 2024.

JHO's research is supported by OpenAI's grant to the 'AI Progress through the Lens of Predictable AI Ecosystems' programme, which is based at the Leverhulme Centre for the Future of Intelligence at the University of Cambridge.

Appendix A. Extended formalisation for complex AI ecosystems

In the main paper, we presented a formal framework that allowed us to precisely define and quantify predictability. However, for simplicity, our framework made key assumptions in only accounting for a single AI system, user and instance at once. We also only considered AI systems to interact with instances through atomic actions. Here, we present an extension of our framework, both considering complete ecosystems of multiple AI systems, users, and task instances, as well as accounting for past information and longer time horizons.

We work with problem instances i, (AI) systems s, (human) users u and system outputs o interacting in particular situations that we call AI ecosystems. We denote the sets of instances $i \in \mathsf{I}$, systems $s \in \mathsf{S}$, users $u \in \mathsf{U}$, and outputs $o \in \mathsf{O}$. Elements of these sets are related by a relation set R, which denotes which instances, systems and users interact to result in particular outputs.

An AI ecosystem at time t is a tuple $\mathsf{E}_t := \langle \mathsf{I}_t, \mathsf{S}_t, \mathsf{U}_t, \mathsf{R}_t \rangle$, specifying the sets of instances, systems and users that are happening at t, related by R_t. A distribution of ecosystems at time t is denoted by $\mathcal{E}_t$. Note that O_t is the set of outputs produced at time t, which we keep separately for convenience, as each of them is simply the result of a system operating on an instance and user: $o_t = s_t(i_t, u_t)$, where $\langle s_t, i_t, u_t \rangle \in \mathsf{R}_t$. We denote by V_t a random variable of a metric of validity at time t, representing how good (correct, safe, etc.) the outcome is produced at time t. $\mathsf{V}_{<\mathsf{t}}$ is used to denote the sequence of validity indicators up to time t. Similarly, $\mathsf{O}_{<\mathsf{t}}$ denotes the sequence of outputs up to time t. The sequence of ecosystems up to and including time t is expressed by $\mathsf{E}_{\leq t}$. In practical scenarios, the full sequence of interactions between instances, AI systems and users may be required to accurately model validity, rather than just the most recent values at time-step t. We therefore rely on $\mathsf{H}_{\leq t} := \langle \mathsf{E}_{\leq t}, \mathsf{O}_{<t}, \mathsf{V}_{<t} \rangle$, the complete sequence history of ecosystems up to and including time t, as well as the observed outputs and validity indicators. This sequence is distributed according to $\mathcal{H}_{\leq t}$, capturing a first kind of stochasticity: the behaviour of the AI systems and the users in the ecosystem.

We denote the probability density function for V_{t+h} (or probability mass function if V_{t+h} is a discrete distribution) given a history of ecosystems as $p(V_{t+h} \mid \mathsf{H}_{\leq t}) = p(V_{t+h} \mid \mathsf{E}_{\leq t}, \mathsf{O}_{<t}, \mathsf{V}_{<t})$ with $h \geq 0$ being the future (or prediction) horizon. This can represent a second type of stochasticity originating from the validity indicator (even for the same history), especially when this validity is reported or assessed by humans. If this possible second source of stochasticity, p, is non-entropic and always assigns the same validity to the same history, the ecosystem can have the first kind of stochasticity in the systems and users. In general, the expected validity can be decomposed into an expression (right) that shows these two sources of stochasticity (on the history and on the validity indicator):

$$\mathbb{V}(p, \mathcal{H}_{\leq t}) := \mathop{\mathbb{E}}_{\mathsf{H}_{\leq t} \sim \mathcal{H}_{\leq t}} [V_{t+h} \mid \mathsf{H}_{\leq t}] = \mathop{\mathbb{E}}_{\mathsf{H}_{\leq t} \sim \mathcal{H}_{\leq t}} \left[\int_v v \cdot p(v \mid \mathsf{H}_{\leq t}) dv \right] \tag{A.1}$$

Note that Eq. 1 in the main text of the paper is a special case of A.1, where the ecosystem history consists of a single time-step over a single AI system, user, and instance.

Likewise, we can extend our notion of unpredictability to the more general case. As in Section 3, we define a family of predictors $\mathcal{F}_b$ bounded on cost or budget b. Once this family is fixed, and relative to it, we can define the unpredictability $\mathbb{Q}$ for a distribution of AI ecosystem histories $\mathcal{H}_{\leq t}$ at time t with prediction horizon h as:

$$\mathbb{Q}(p, \mathcal{H}_t, \mathcal{F}_b) := \min_{\hat{p} \in \mathcal{F}_b} \mathop{\mathbb{E}}_{\substack{\mathsf{H}_{\leq t} \sim \mathcal{H}_{\leq t} \\ v \sim p(V_{t+h} \mid \mathsf{H}_t)}} S(\hat{p}(V_{t+h} \mid H_{\leq t}), v) \tag{A.2}$$

with S being a function that evaluates the probabilistic predictions against the observed validity values, such as any well-defined proper scoring rule (PSR). Returning to the self-driving car example and Fig. 1, if we set $\mathcal{F}_b$ as the family of logistic functions with b features, then as with Eq. 2, we can see that the unpredictability of A using one feature (i.e., $\mathbb{Q}(p_A, \mathcal{H}_{\leq t}, \mathcal{F}_1)$) is low. However, now

we are able to also model an entire fleet of self driving cars serving multiple users over multiple trips, and assess the validity of the trips as they occur (i.e., given that the past five trips have been successful for a car, how likely is it that it will continue to be so?).

Note that given the family $\mathcal{F}$ of all computable functions, if p has zero entropy (the ecosystem would be deterministic), then we could approach 0 unpredictability with enough data, time and other resources. In practice, finding a perfect $\hat{p}$ for some arbitrary p would be intractable. For instance, for some machine learning architectural families, the budget b would be set on some computation limits assuming access to the history of ecosystems, outputs and validity values before t as training set. However, even with unlimited computational resources, if the underlying distribution p is stochastic, the loss may not be zero. This is due to aleatoric uncertainty, which is the inherent unpredictability of a system or process. For instance, suppose that both the estimated probability distribution $\hat{p}$ and the true probability distribution p consistently assign the probability of an event $\in \{0, 1\}$ to be 0.7 (e.g., a biased coin whose head and tail have a probability of 0.7 and 0.3, respectively). Then, with this best possible predictor, the Brier score and cross-entropy loss are $0.7 \cdot (0.7-1)^2 + 0.3 \cdot (0.7-0)^2 = 0.21$ and $-(0.7\log(0.7) + 0.3\log(0.3)) \approx 0.61$, respectively, instead of 0.

Of course, when the ecosystem's AI models are maximally optimal (i.e., they never fail nor produce invalid outputs), then the unpredictability of the ecosystem disappears since using such systems always yields the maximum validity V. Formally, if $\forall \mathsf{H} \in \mathcal{H} : p(V = v_{\max} \mid \mathsf{H}) = 1$ then any family $\mathcal{F}$ that contains the constant predictor $\hat{p} = v_{\max}$ makes $\mathbb{Q} = 0$. This is a generalisation of the third row in Table 2.

The opposite extreme is the worst-case scenario where AI models always produce the worst possible outcomes (i.e., minimal or zero validity): $\forall \mathsf{H} \in \mathcal{H} : p(V = v_{\min} \mid \mathsf{H}) = 1$. Similarly, any family $\mathcal{F}$ that contains the constant predictor $\hat{p} = v_{\min}$ makes $\mathbb{Q} = 0$. Again, this is a generalisation of the fourth row in Table 2. It is because of this pessimal case that, for predictable AI, we want to find a Pareto frontier that balances minimising $\mathbb{Q}$ while maximising expected validity $\mathbb{V}$, or to optimise for some metrics that combine high validity and low rejection using $\hat{p}$, such as the area under the accuracy-rejection curve [148,172].

With the extended formulation in this appendix, we can define the *validity envelope* as the largest subset of the distribution $\mathcal{E}^{\omega,\sigma} \subset \mathcal{E}$, where expected validity is no smaller than ω, predicted with a loss of at most σ:

$$\mathbb{V}(\hat{p}, \mathcal{E}^{\omega,\sigma}) \geq \omega \;\wedge\; \left[\mathop{\mathbb{E}}_{\substack{\langle i,s,u\rangle \sim \mathcal{E}^{\omega,\sigma} \\ v \sim p(V \mid \langle s,i,u\rangle)}} S(\hat{p}(V \mid \langle i, s, u\rangle), v) \right] \leq \sigma \tag{A.3}$$

The importance of the validity envelope is that we can determine where to operate according to the constraints about fairness, reward, scores, energy, response time, etc., through reject rules or other assurance mechanisms.

References

[1] M. Dehghani, Y. Tay, A.A. Gritsenko, Z. Zhao, N. Houlsby, F. Diaz, D. Metzler, O. Vinyals, The benchmark lottery, (2021) arXiv preprint arXiv:2107.07002.
[2] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, K.-R. Müller, Unmasking clever hans predictors and assessing what machines really learn, Nat. Commun. 10 (1) (2019) 1096.
[3] C. Vásquez-Venegas, C. Wu, S. Sundar, R. Prôa, F.J. Beloy, J.R. Medina, M. McNichol, K. Parvataneni, N. Kurtzman, F. Mirshawka, et al., Detecting and mitigating the clever hans effect in medical imaging: a scoping review, J. Imaging Informat. Med. (2024) 1–17.
[4] J. Skalse, N. Howe, D. Krasheninnikov, D. Krueger, Defining and characterizing reward gaming, Adv. Neural Inf. Process. Syst. 35 (2022) 9460–9471.
[5] A. Bondarenko, D. Volk, D. Volkov, J. Ladish, Demonstrating specification gaming in reasoning models, 2025, https://arxiv.org/abs/2502.13295.
[6] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, et al., Emergent abilities of large language models, (2022) arXiv preprint arXiv:2206.07682.
[7] L. Berti, F. Giorgi, G. Kasneci, Emergent abilities in large language models: a survey, (2025) arXiv preprint arXiv:2503.05788.
[8] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y.J. Bang, A. Madotto, P. Fung, Survey of hallucination in natural language generation, ACM Comp. Surv. 55 (12) (2023) 1–38.
[9] Y. Zhang, S. Li, C. Qian, J. Liu, P. Yu, C. Han, Y.R. Fung, K. McKeown, C. Zhai, M. Li, et al., The law of knowledge overshadowing: towards understanding, predicting, and preventing LLM hallucination, (2025) arXiv preprint arXiv:2502.16143.
[10] A. Tamkin, M. Brundage, J. Clark, D. Ganguli, Understanding the capabilities, limitations, and societal impact of large language models, (2021) arXiv preprint arXiv:2102.02503.
[11] L. Zhou, W. Schellaert, F. Martínez-Plumed, Y. Moros-Daval, C. Ferri, J. Hernández-Orallo, Larger and more instructable language models become less reliable, Nature 634 (2024) 61–68. https://doi.org/10.1038/s41586-024-07930-y
[12] R.L. Thomas, D. Uminsky, Reliance on metrics is a fundamental challenge for AI, Patterns 3 (5) (2022) 100476. https://doi.org/10.1016/j.patter.2022.100476
[13] M. Eriksson, E. Purificato, A. Noroozian, J. Vinagre, G. Chaslot, E. Gomez, D. Fernandez-Llorca, Can we trust AI benchmarks? an interdisciplinary review of current issues in AI evaluation, (2025) arXiv preprint arXiv:2502.06559.
[14] J. Fodor, Line goes up? Inherent limitations of benchmarks for evaluating large language models, (2025) arXiv preprint arXiv:2502.14318.
[15] R. Kocielnik, S. Amershi, P.N. Bennett, Will you accept an imperfect AI? exploring designs for adjusting end-user expectations of AI systems, in: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 2019, pp. 1–14.
[16] D. Schuster, Abstaining machine learning: philosophical considerations, AI & Society (2025) 1–21.
[17] D. Hersman, Safety at Waymo | Waymo and the weather, 2019, Accessed on September 13, 2024, https://waymo.com/blog/2019/08/waymo-and-weather/.
[18] S. Zang, M. Ding, D. Smith, P. Tyler, T. Rakotoarivelo, M.A. Kaafar, The impact of adverse weather conditions on autonomous vehicles: how rain, snow, fog, and hail affect the performance of a self-driving car, IEEE Veh. Technol. Mag. 14 (2) (2019) 103–111.
[19] J. Betz, M. Lutwitzi, S. Peters, A new taxonomy for automated driving: structuring applications based on their operational design domain, level of automation and automation readiness, (2024) arXiv preprint arXiv:2404.17044.
[20] I. Hauhio, A. Kantosalo, S. Linkola, H. Toivonen, The spectrum of unpredictability and its relation to creative autonomy, in: International Conference on Computational Creativity, Association for Computational Creativity (ACC), 2023, pp. 148–152.
[21] R. Burnell, J. Burden, D. Rutar, K. Voudouris, L. Cheke, J. Hernández-Orallo, Not a number: identifying instance features for capability-oriented evaluation, in: IJCAI, 2022, pp. 2827–2835.
[22] M. Taddeo, M. Ziosi, A. Tsamados, L. Gilli, S. Kurapati, Artificial intelligence for national security: the predictability problem, Cent. Digit. Ethics Res. Pap. No. (2022).
[23] N. Carlini, A GPT-4 capability forecasting challenge, 2024, (https://nicholas.carlini.com/writing/llm-forecast/question/Capital-of-Paris). Accessed: 2024-09-08.

[24] S. Kandul, V. Micheli, J. Beck, T. Burri, F. Fleuret, M. Kneer, M. Christen, Human control redressed: comparing AI and human predictability in a real-effort task, Comp. Hum. Behav. Rep. 10 (2023) 100290. https://doi.org/10.1016/j.chbr.2023.100290
[25] HLEG-AI-EC, Ethics guidelines for trustworty artificial intelligence, Technical Report, High-Level Expert Group on Artificial Intelligence, European Commission, 2019.
[26] EU-AI-Act, Eu artificial intelligence act, , 2024, (Regulation (EU) 2024/1689, Official Journal). Interinstitutional File: 2021/0106(COD), https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32024R1689.
[27] S. Passi, M. Vorvoreanu, Overreliance on AI literature review, Microsoft Res. (2022).
[28] ISO/IEC, Information technology artificial intelligence artificial intelligence concepts and terminology, Draft International Standard 22989, International Organization for Standardization, Geneva, Switzerland, 2022. Definition 3.4.7, https://www.iso.org/obp/ui/fr/#iso:std:iso-iec:22989:dis:ed-1:v1:en.
[29] D.F. Llorca, V. Charisi, R. Hamon, I. Sánchez, E. Gómez, Liability regimes in the age of AI: a use-case driven analysis of the burden of proof, J. Artif. Intell. Res. 76 (2023) 613–644.
[30] B. Green, The flaws of policies requiring human oversight of government algorithms, Comp. Law Secur. Rev. 45 (2022) 105681.
[31] R. Koulu, Human control over automation: EU policy and AI ethics, Eur. J. Leg. Stud. 12 (2020) 9.
[32] J. Beck, T. Burri, M. Christen, F. Fleuret, S. Kandul, M. Kneer, V. Micheli, Human control redressed: comparing AI-to-human vs. human-to-human predictability in a real-effort task, Hum.-To-Hum. Predict. Real-Effort Task (2023).
[33] I. Gabriel, Artificial intelligence, values, and alignment, Minds Mach. 30 (3) (2020) 411–437.
[34] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, D. Mané, Concrete problems in AI safety, (2016) arXiv preprint arXiv:1606.06565.
[35] M. Sharma, M. Tong, J. Mu, J. Wei, J. Kruthoff, S. Goodfriend, E. Ong, A. Peng, R. Agarwal, C. Anil, et al., Constitutional classifiers: defending against universal jailbreaks across thousands of hours of red teaming, (2025) arXiv preprint arXiv:2501.18837.
[36] I. Rahwan, M. Cebrian, N. Obradovich, J. Bongard, J.-F. Bonnefon, C. Breazeal, J.W. Crandall, N.A. Christakis, I.D. Couzin, M.O. Jackson, et al., Machine behaviour, Nature 568 (7753) (2019) 477–486.
[37] R.V. Yampolskiy, Unpredictability of AI, (2019) arXiv preprint arXiv:1905.13053.
[38] R.V. Yampolskiy, AI: unexplainable, unpredictable, uncontrollable, CRC Press, 2024.
[39] D. Ganguli, D. Hernandez, L. Lovitt, A. Askell, Y. Bai, A. Chen, T. Conerly, N. Dassarma, D. Drain, N. Elhage, et al., Predictability and surprise in large generative models, in: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, 2022, pp. 1747–1764.
[40] E. Hüllermeier, W. Waegeman, Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods, Mach. Learn. 110 (3) (2021) 457–506.
[41] P.J. Roache, Verification and validation in computational science and engineering, 895, Hermosa Albuquerque, NM, 1998.
[42] J.M. Zhang, M. Harman, L. Ma, Y. Liu, Machine learning testing: survey, landscapes and horizons, IEEE Transp. Softw. Eng. 48 (1) (2020) 1–36.
[43] J.P. Miller, R. Taori, A. Raghunathan, S. Sagawa, P.W. Koh, V. Shankar, P. Liang, Y. Carmon, L. Schmidt, Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization, in: International Conference on Machine Learning, PMLR, 2021, pp. 7721–7735.
[44] B. Benifei, I.-D. Tudorache, Proposal for a regulation of the European Parliament and of the Council on harmonised rules on artificial intelligence (Artificial Intelligence Act), Technical Report, Tech. rep., Committee on the Internal Market and Consumer Protection …, 2023. https://www.europarl.europa.eu/meetdocs/2014_2019/plmrep/COMMITTEES/CJ40/DV/2023/05-11/ConsolidatedCA_IMCOLIBE_AI_ACT_EN.pdf.
[45] S. Armstrong, K. Sotala, Ó.h. Seán S., The errors, insights and lessons of famous AI predictions–and what they mean for the future, J. Exp. Theor. Artif. Intell. 26 (3) (2014) 317–342.
[46] R. Gruetzemacher, F.E. Dorner, N. Bernaola-Alvarez, C. Giattino, D. Manheim, Forecasting AI progress: a research agenda, Technol. Forecast. Soc. Change 170 (2021) 120909.
[47] C.B. Frey, M.A. Osborne, The future of employment: how susceptible are jobs to computerisation?, Technol. Forecast. Soc. Change 114 (2017) 254–280.
[48] S. Tolan, A. Pesole, F. Martínez-Plumed, E. Fernández-Macías, J. Hernández-Orallo, E. Gómez, Measuring the occupational impact of AI: tasks, cognitive abilities and AI benchmarks, J. Artif. Intell. Res. 71 (2021) 191–236.
[49] T. Eloundou, S. Manning, P. Mishkin, D. Rock, Gpts are GPTs: an early look at the labor market impact potential of large language models, (2023) arXiv preprint arXiv:2303.10130.
[50] M. Staneva, S. Elliott, Measuring the impact of artificial intelligence and robotics on the workplace, in: New Digital Work: digital Sovereignty at the Workplace, Springer, 2023, pp. 16–30.
[51] I. Ong, A. Almahairi, V. Wu, W.-L. Chiang, T. Wu, J.E. Gonzalez, M.W. Kadous, I. Stoica, Routellm: learning to route llms with preference data, (2024) arXiv preprint arXiv:2406.18665.
[52] L. Zhou, F. Martínez-Plumed, J. Hernández-Orallo, C. Ferri, W. Schellaert, Reject before you run: small assessors anticipate big language models, in: Ebem@Ijcai, 2022.
[53] R. OpenAI, GPT-4 technical report, (2023). https://doi.org/10.48550/arXiv.2303.08774
[54] Microsoft, Prompt shields, 2024, Accessed: 2024-09-08, https://learn.microsoft.com/en-us/azure/ai-services/content-safety/concepts/jailbreak-detection.
[55] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, et al., The llama 3 herd of models, (2024) arXiv preprint arXiv:2407.21783.
[56] G. Bansal, B. Nushi, E. Kamar, W.S. Lasecki, D.S. Weld, E. Horvitz, Beyond accuracy: the role of mental models in human-AI team performance, in: Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, 7, 2019, pp. 2–11.
[57] Y. Zhao, L. Ju, J. Hernández-Orallo, Team formation through an assessor: choosing MARL agents in pursuit–evasion games, Complex Intell. Syst. (2024) 1–20.
[58] R. Trivedi, A. Khan, J.C.L. Hammond, E.A. Duéñez-Guzmán, D. Chakraborty, J.P. Agapiou, J. Matyas, S. Vezhnevets, B. Pásztor, Y. Ao, O.G. Younis, J. Huang, B. Swain, H. Qin, M. Deng, Z. Deng, U. Erdoğanaras, Y. Zhao, M. Tesic, N. Jaques, J.N. Foerster, V. Conitzer, J. Hernandez-Orallo, D. Hadfield-Menell, J.Z. Leibo, Melting pot contest: charting the future of generalized cooperative intelligence, NeurIPS (2024).
[59] J. Kaplan, S. McCandlish, T. Henighan, T.B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, D. Amodei, Scaling laws for neural language models, (2020) arXiv preprint arXiv:2001.08361.
[60] J.C. Kaminski, C. Hopp, Predicting outcomes in crowdfunding campaigns with textual, visual, and linguistic signals, Small Bus. Econ. 55 (3) (2020) 627–649.
[61] S. Gu, Several questions of visual generation in 2024, (2024) arXiv preprint arXiv:2407.18290.
[62] J. Becker, N. Rush, B. Barnes, D. Rein, Measuring the impact of early-2025 AI on experienced open-source developer productivity, Model Evaluat. Threat Res. (METR), (2025). https://metr.org/Early_2025_AI_Experienced_OS_Devs_Study.pdf
[63] X. Hou, Y. Zhao, Y. Liu, Z. Yang, K. Wang, L. Li, X. Luo, D. Lo, J. Grundy, H. Wang, Large language models for software engineering: a systematic literature review, ACM Transp. Softw. Eng. Methodol. (2024). Just Accepted, https://doi.org/10.1145/3695988
[64] L. Peracchio, G. Nicora, E. Parimbelli, T.M. Buonocore, R. Bergamaschi, E. Tavazzi, A. Dagliati, R. Bellazzi, Evaluation of predictive reliability to foster trust in artificial intelligence. a case study in multiple sclerosis, (2024) arXiv preprint arXiv:2402.17554.
[65] J. Feng, A. Subbaswamy, A. Gossmann, H. Singh, B. Sahiner, M.-O. Kim, G.A. Pennello, N. Petrick, R. Pirracchio, F. Xia, Designing monitoring strategies for deployed machine learning algorithms: navigating performativity through a causal lens, in: Causal Learning and Reasoning, PMLR, 2024, pp. 587–608.
[66] C. Mougan, D.S. Nielsen, Monitoring model deterioration with explainable uncertainty estimation via non-parametric bootstrap, in: Proceedings of the AAAI Conference on Artificial Intelligence, 37, 2023, pp. 15037–15045.
[67] X. Wei, X. Chu, J. Geng, Y. Wang, P. Wang, H. Wang, C. Wang, L. Lei, Societal impacts of chatbot and mitigation strategies for negative impacts: a large-scale qualitative survey of chatGPT users, Technol. Soc. 77 (2024) 102566.
[68] R. Goebel, A. Chander, K. Holzinger, F. Lecue, Z. Akata, S. Stumpf, P. Kieseberg, A. Holzinger, Explainable AI: the new 42?, in: International Cross-Domain Conference for Machine Learning and Knowledge Extraction, Springer, 2018, pp. 295–303.
[69] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, G.-Z. Yang, XAI-explainable artificial intelligence, Sci. Rob. 4 (37) (2019) eaay7120.
[70] T. Miller, Explanation in artificial intelligence: insights from the social sciences, Artif. Intell. 267 (2019) 1–38.

[71] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM Comp. Surv. (CSUR) 51 (5) (2018) 1–42.
[72] C. Molnar, Interpretable machine learning, Lulu. com, 2020.
[73] L. Bereska, E. Gavves, Mechanistic interpretability for AI safetya review, (2024) arXiv preprint arXiv:2404.14082.
[74] C. Giraud-Carrier, R. Vilalta, P. Brazdil, Introduction to the special issue on meta-learning, Mach. Learn. 54 (2004) 187–193.
[75] J. Vanschoren, Meta-learning: a survey, (2018) arXiv preprint arXiv:1810.03548.
[76] A. Bella, C. Ferri, J. Hernández-Orallo, M.J. Ramírez-Quintana, Calibration of machine learning models, in: Handbook of Research on Machine Learning Applications and Trends: algorithms, methods, and techniques, IGI Global, 2010, pp. 128–146.
[77] J. Nixon, M.W. Dusenberry, L. Zhang, G. Jerfel, D. Tran, Measuring calibration in deep learning, in: CVPR Workshops, 2, 2019.
[78] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U.R. Acharya, et al., A review of uncertainty quantification in deep learning: techniques, applications and challenges, Inf. Fusion 76 (2021) 243–297.
[79] J. Gawlikowski, C.R.N. Tassi, M. Ali, J. Lee, M. Humt, J. Feng, A. Kruspe, R. Triebel, P. Jung, R. Roscher, et al., A survey of uncertainty in deep neural networks, Artif. Intell. Rev. 56 (Suppl 1) (2023) 1513–1589.
[80] B. Schölkopf, F. Locatello, S. Bauer, N.R. Ke, N. Kalchbrenner, A. Goyal, Y. Bengio, Toward causal representation learning, Proc. IEEE 109 (5) (2021) 612–634.
[81] J. Peters, D. Janzing, B. Schölkopf, Elements of causal inference: foundations and learning algorithms, The MIT Press, 2017.
[82] D. Pessach, E. Shmueli, A review on fairness in machine learning, ACM Comp. Surv. 55 (3) (2022) 1–44.
[83] D. Dalrymple, J. Skalse, Y. Bengio, S. Russell, M. Tegmark, S. Seshia, S. Omohundro, C. Szegedy, B. Goldhaber, N. Ammann, et al., Towards guaranteed safe AI: a framework for ensuring robust and reliable AI systems, (2024) arXiv preprint arXiv:2405.06624.
[84] L.G. Valiant, A theory of the learnable, Commun. ACM 27 (11) (1984) 1134–1142.
[85] G. Fink, M. Bishop, Property-based testing: a new approach to testing for assurance, ACM SIGSOFT Softw. Eng. Notes 22 (4) (1997) 74–80.
[86] M.T. Ribeiro, T. Wu, C. Guestrin, S. Singh, Beyond accuracy: behavioral testing of NLP models with CheckList, (2020) arXiv preprint arXiv:2005.04118.
[87] H. Lightman, V. Kosaraju, Y. Burda, H. Edwards, B. Baker, T. Lee, J. Leike, J. Schulman, I. Sutskever, K. Cobbe, Let's verify step by step, (2023) arXiv preprint arXiv:2305.20050.
[88] E. Jones, M. Tong, J. Mu, M. Mahfoud, J. Leike, R. Grosse, J. Kaplan, W. Fithian, E. Perez, M. Sharma, Forecasting rare language model behaviors, (2025) arXiv preprint arXiv:2502.16797.
[89] J. Sevilla, T. Besiroglu, B. Cottier, J. You, E. Roldán, P. Villalobos, E. Erdil, Can AI scaling continue through 2030?, 2024, Accessed: 2024-10-04, https://epochai.org/blog/can-ai-scaling-continue-through-2030.
[90] A.I. Epoch, Data on notable AI models, 2024, Accessed: 2024-10-04, https://epochai.org/data/notable-ai-models.
[91] D. Rein, J. Becker, A. Deng, S. Nix, C. Canal, D. O'Connel, P. Arnott, R. Bloom, T. Broadley, K. Garcia, et al., HCAST: human-calibrated autonomy software tasks, (2025) arXiv preprint arXiv:2503.17354.
[92] T. Kwa, B. West, J. Becker, A. Deng, K. Garcia, M. Hasin, S. Jawhar, M. Kinniment, N. Rush, S. Von Arx, et al., Measuring AI ability to complete long tasks, (2025) arXiv preprint arXiv:2503.14499.
[93] S.E. Middleton, E. Letouzé, A. Hossaini, A. Chapman, Trust, regulation, and human-in-the-loop AI: within the european region, Commun. ACM 65 (4) (2022) 64–68.
[94] B. Nushi, E. Kamar, E. Horvitz, Towards accountable AI: hybrid human-machine analyses for characterizing system failure, in: Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, 6, 2018, pp. 126–135.
[95] B.M. Lake, T.D. Ullman, J.B. Tenenbaum, S.J. Gershman, Building machines that learn and think like people, Behav. Brain Sci. 40 (2017) e253.
[96] I. Momennejad, A rubric for human-like agents and neuroAI, Phil. Transp. Roy. Soc. B 378 (1869) (2023) 20210446.
[97] E. Brynjolfsson, The turing trap: the promise & peril of human-like artificial intelligence, Daedalus 151 (2) (2022) 272–287.
[98] K. Grace, J. Salvatier, A. Dafoe, B. Zhang, O. Evans, When will AI exceed human performance? evidence from AI experts, J. Artif. Intell. Res. 62 (2018) 729–754.
[99] R. Gruetzemacher, D. Paradice, K.B. Lee, Forecasting transformative AI: an expert survey, (2019) arXiv preprint arXiv:1901.08579.
[100] J. Steinhardt, What will GPT-2030 look like?, 2023, (Bounded Regret). https://bounded-regret.ghost.io/what-will-gpt-2030-look-like/.
[101] E. Karger, J. Rosenberg, Z. Jacobs, M. Hickman, R. Hadshar, K. Gamin, T. Smith, B. Williams, T. McCaslin, P.E. Tetlock, Forecasting existential risks evidence from a long-run forecasting tournament, FRI Work. Pap. (2023).
[102] P. Mühlbacher, P. Scoblic, Exploring Metaculus's AI track record, 2024, (Metaculus Journal). https://www.metaculus.com/notebooks/16708/exploring-metaculuss-ai-track-record/.
[103] K.J. Arrow, R. Forsythe, M. Gorham, R. Hahn, R. Hanson, J.O. Ledyard, S. Levmore, R. Litan, P. Milgrom, F.D. Nelson, et al., The promise of prediction markets, 2008.
[104] M.A. Burgman, Trusting judgements: how to get the best out of experts, Cambridge University Press, 2016.
[105] Z. Jiang, J. Araki, H. Ding, G. Neubig, How can we know when language models know? on the calibration of language models for question answering, Transp. Assoc. Comput. Linguist. 9 (2021) 962–977.
[106] Y. Xiao, P.P. Liang, U. Bhatt, W. Neiswanger, R. Salakhutdinov, L.-P. Morency, Uncertainty quantification with pre-trained language models: a large-scale empirical analysis, (2022) arXiv preprint arXiv:2210.04714.
[107] D. Krueger, T. Maharaj, J. Leike, Hidden incentives for auto-induced distributional shift, (2020) arXiv preprint arXiv:2009.09153.
[108] D. Hendrycks, N. Carlini, J. Schulman, J. Steinhardt, Unsolved problems in ml safety, (2021) arXiv preprint arXiv:2109.13916.
[109] R. Kamoi, Y. Zhang, N. Zhang, J. Han, R. Zhang, When can LLMs actually correct their own mistakes? a critical survey of self-correction of LLMs, Transp. Assoc. Comput. Linguist. 12 (2024) 1417–1440.
[110] J. Hernández-Orallo, W. Schellaert, F. Martínez-Plumed, Training on the test set: mapping the system-problem space in AI, in: Proceedings of the AAAI Conference on Artificial Intelligence, 36, 2022, pp. 12256–12261.
[111] W. Schellaert, F. Martínez-Plumed, J. Hernández-Orallo, Analysing the predictability of language model performance, ACM Transp. Intell. Syst. Technol. 16 (2) (2025) 1–26.
[112] W. Schellaert, The evaluation of artificial intelligence as a prediction problem, Ph.D. thesis, PhD thesis, Universitat Politecnica de Valencia, 2025.
[113] J. Burden, K. Voudouris, R. Burnell, D. Rutar, L. Cheke, J. Hernández-Orallo, Inferring capabilities from task performance with Bayesian triangulation, (2023) arXiv preprint arXiv:2309.11975.
[114] J. Burden, L. Cheke, J. Hernandez-Orallo, M. Tešić, K. Voudouris, Measurement layouts for capability-oriented AI evaluation, 2024.
[115] L. Zhou, L. Pacchiardi, F. Martínez-Plumed, K.M. Collins, Y. Moros-Daval, S. Zhang, Q. Zhao, Y. Huang, L. Sun, J.E. Prunty, et al., General scales unlock AI evaluation with explanatory and predictive power, (2025) arXiv preprint arXiv:2503.06378.
[116] P. Shojaee, I. Mirzadeh, K. Alizadeh, M. Horton, S. Bengio, M. Farajtabar, The illusion of thinking: understanding the strengths and limitations of reasoning models via the lens of problem complexity, (2025) arXiv preprint arXiv:2506.06941.
[117] J. Hernández-Orallo, Evaluation in artificial intelligence: from task-oriented to ability-oriented measurement, Artif. Intell. Rev. 48 (2017) 397–447.
[118] J. Hernández-Orallo, The measure of all minds: evaluating natural and artificial intelligence, Cambridge University Press, 2017.
[119] R.B.C. Prudêncio, A.C. Lorena, T. Silva-Filho, P. Drapal, M.G. Valeriano, Assessor models for explaining instance hardness in classification problems, in: 2024 International Joint Conference on Neural Networks (IJCNN), IEEE, 2024, pp. 1–8.
[120] D. Ashok, J. May, Language models can predict their own behavior, (2025) arXiv preprint arXiv:2502.13329.
[121] R. Burnell, W. Schellaert, J. Burden, T.D. Ullman, F. Martinez-Plumed, J.B. Tenenbaum, D. Rutar, L.G. Cheke, J. Sohl-Dickstein, M. Mitchell, et al., Rethink reporting of evaluation results in AI, Sci. 380 (6641) (2023) 136–138.
[122] P.F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, D. Amodei, Deep reinforcement learning from human preferences, Adv. Neural Inf. Process. Syst. 30 (2017).

[123] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al., Training language models to follow instructions with human feedback, Adv. Neural Inf. Process. Syst. 35 (2022) 27730–27744.
[124] A. Glaese, N. McAleese, M. Trębacz, J. Aslanides, V. Firoiu, T. Ewalds, M. Rauh, L. Weidinger, M. Chadwick, P. Thacker, et al., Improving alignment of dialogue agents via targeted human judgements, (2022) arXiv preprint arXiv:2209.14375.
[125] Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon, et al., Constitutional AI: harmlessness from AI feedback, (2022) arXiv preprint arXiv:2212.08073.
[126] Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan, et al., Training a helpful and harmless assistant with reinforcement learning from human feedback, (2022) arXiv preprint arXiv:2204.05862.
[127] A. Chan, Evaluating predictions of model behaviour, 2024, (Centre for the Governance of AI. https://www.governance.ai/analysis/evaluating-predictions-of-model-behaviour).
[128] G. Zhang, F.E. Dorner, M. Hardt, How benchmark prediction from fewer data misses the mark, (2025) arXiv preprint arXiv:2506.07673.
[129] T. Tamura, T. Yano, M. Enomoto, M. Oyamada, Can a crow hatch a falcon? lineage matters in predicting large language model performance, (2025) arXiv preprint arXiv:2504.19811.
[130] D. Hernandez, J. Kaplan, T. Henighan, S. McCandlish, Scaling laws for transfer, (2021) arXiv preprint arXiv:2102.01293.
[131] D. Owen, How predictable is language model benchmark performance?, (2024) arXiv preprint arXiv:2401.04757.
[132] F.M. Polo, S. Somerstep, L. Choshen, Y. Sun, M. Yurochkin, Sloth: scaling laws for LLM skills to predict multi-benchmark performance across families, (2025) arXiv preprint arXiv:2412.06540.
[133] F. Martinez-Plumed, J. Hernandez-Orallo, E.G. Gutierrez, AI watch: methodology to monitor the evolution of AI technologies, Technical Report, Joint Research Centre (Seville site), 2020.
[134] F. Martínez-Plumed, J. Hernández-Orallo, E. Gómez, Tracking AI: the capability is (not) near, in: Ecai 2020, IOS Press, 2020, pp. 2915–2916.
[135] D. Zhang, N. Maslej, E. Brynjolfsson, J. Etchemendy, T. Lyons, J. Manyika, H. Ngo, J.C. Niebles, M. Sellitto, E. Sakhaee, Y. Shoham, J. Clark, R. Perrault, The AI Index 2022 annual report, 2022. arxiv:2205.03468
[136] S. Kadavath, T. Conerly, A. Askell, T. Henighan, D. Drain, E. Perez, N. Schiefer, Z. Hatfield-Dodds, N. DasSarma, E. Tran-Johnson, et al., Language models (mostly) know what they know, (2022) arXiv preprint arXiv:2207.05221.
[137] M. Crosby, B. Beyret, J. Hernandez-Orallo, L. Cheke, M. Halina, M. Shanahan, Translating from animal cognition to AI, in: NeurIPS Workshop on Biological and Artificial Reinforcement Learning, 2019.
[138] M. Crosby, B. Beyret, M. Shanahan, J. Hernández-Orallo, L. Cheke, M. Halina, The animal-AI testbed and competition, in: H.J. Escalante, R. Hadsell (Eds.), Proceedings of the NeurIPS 2019 Competition and Demonstration Track, 123 of *Proceedings of Machine Learning Research*, PMLR, 2020, pp. 164–176. https://proceedings.mlr.press/v123/crosby20a.html.
[139] F. Brahman, S. Kumar, V. Balachandran, P. Dasigi, V. Pyatkin, A. Ravichander, S. Wiegreffe, N. Dziri, K. Chandu, J. Hessel, et al., The art of saying no: contextual noncompliance in language models, (2024) arXiv preprint arXiv:2407.12043.
[140] B. Dreyfuss, R. Raux, Human learning about AI, 2025.2406.05408 https://arxiv.org/abs/2406.05408.
[141] L. Pacchiardi, K. Voudouris, B. Slater, F. Martínez-Plumed, J. Hernández-Orallo, L. Zhou, W. Schellaert, PredictaBoard: benchmarking LLM score predictability, (2025) arXiv preprint arXiv:2502.14445.
[142] W. Schellaert, R. Hamon, F. Martínez-Plumed, J. Hernandez-Orallo, A proposal for scaling the scaling laws, in: Proceedings of the First Edition of the Workshop on the Scaling Behavior of Large Language Models (SCALE-LLM 2024), 2024, pp. 1–8.
[143] T. Zhong, Z. Liu, Y. Pan, Y. Zhang, Y. Zhou, S. Liang, Z. Wu, Y. Lyu, P. Shu, X. Yu, et al., Evaluation of OpenAI o1: opportunities and challenges of AGI, (2024) arXiv preprint arXiv:2409.18486.
[144] R. Schaeffer, H. Schoelkopf, B. Miranda, G. Mukobi, V. Madan, A. Ibrahim, H. Bradley, S. Biderman, S. Koyejo, Why has predicting downstream capabilities of frontier AI models with scale remained elusive?, ICML 2024 Workshop NextGenAISafety homepage, arXiv preprint arXiv:2406.04391 (2024).
[145] G. Pimpale, A. Højmark, J. Scheurer, M. Hobbhahn, Forecasting frontier language model agent capabilities, (2025) arXiv preprint arXiv:2502.15850.
[146] K.F. Pilz, L. Heim, N. Brown, Increased compute efficiency and the diffusion of AI capabilities, in: Proceedings of the AAAI Conference on Artificial Intelligence, 39, 2025, pp. 27582–27590.
[147] R. Schaeffer, B. Miranda, S. Koyejo, Are emergent abilities of large language models a mirage?, Adv. Neural Inf. Process. Syst. 36 (2024).
[148] F. Condessa, J. Bioucas-Dias, J. Kovačević, Performance measures for classification systems with rejection, Pattern Recognit 63 (2017) 437–450.
[149] L. Pacchiardi, K. Voudouris, B. Slater, F. Martínez-Plumed, J. Hernández-Orallo, L. Zhou, W. Schellaert, PredictaBoard: benchmarking LLM score predictability, (2025) arXiv preprint arXiv:2502.14445.
[150] A. Askell, Y. Bai, A. Chen, D. Drain, D. Ganguli, T. Henighan, A. Jones, N. Joseph, B. Mann, N. DasSarma, et al., A general language assistant as a laboratory for alignment, (2021) arXiv preprint arXiv:2112.00861.
[151] X. Wang, L. Jiang, J. Hernandez-Orallo, D. Stillwell, L. Sun, F. Luo, X. Xie, Evaluating general-purpose ai with psychometrics, Commun. ACM , to appear, arXiv preprint arXiv:2310.16379 (2025).
[152] P. Liang, R. Bommasani, T. Lee, D. Tsipras, D. Soylu, M. Yasunaga, Y. Zhang, D. Narayanan, Y. Wu, A. Kumar, et al., Holistic evaluation of language models, (2022) arXiv preprint arXiv:2211.09110.
[153] H. Toner, P. Hall, S. McGregor, AI incident database, 2021. https://incidentdatabase.ai/.
[154] F. Gilardi, M. Alizadeh, M. Kubli, ChatGPT outperforms crowd workers for text-annotation tasks, Proc. Natl. Acad. Sci. 120 (30) (2023) e2305016120.
[155] A. Esuli, A. Fabris, A. Moreo, F. Sebastiani, Learning to quantify, Springer Nature, 2023.
[156] T. Schumacher, M. Strohmaier, F. Lemmerich, A comparative evaluation of quantification methods, J. Mach. Learn. Res. 26 (55) (2025) 1–54.
[157] T. Li, A.K. Sahu, A. Talwalkar, V. Smith, Federated learning: challenges, methods, and future directions, IEEE Signal Process. Mag. 37 (3) (2020) 50–60.
[158] D. Manheim, S. Martin, M. Bailey, M. Samin, R. Greutzmacher, The necessity of AI audit standards boards, AI SOCIETY (2025) 1–16.
[159] E. Grunberg, F. Modigliani, The predictability of social events, J. Polit. Econ. 62 (6) (1954) 465–478.
[160] D.L. Stern, V. Orgogozo, Is genetic evolution predictable?, Sci. 323 (5915) (2009) 746–751.
[161] C. Song, Z. Qu, N. Blumm, A.-L. Barabási, Limits of predictability in human mobility, Sci. 327 (5968) (2010) 1018–1021.
[162] S. Conway Morris, Evolution: like any other science it is predictable, Phil. Transp. Roy. Soc. B 365 (1537) (2010) 133–145. https://doi.org/10.1098/rstb.2009.0154
[163] C.T. Kello, G.D.A. Brown, R. Ferrer-i Cancho, J.G. Holden, K. Linkenkaer-Hansen, T. Rhodes, G.C. Van Orden, Scaling laws in cognitive sciences, Trends Cogn. Sci. (Regul. Ed.) 14 (5) (2010) 223–232.
[164] M. Kosinski, D. Stillwell, T. Graepel, Private traits and attributes are predictable from digital records of human behavior, Proc. Natl. Acad. Sci. 110 (15) (2013) 5802–5805.
[165] J. Svegliato, K.H. Wray, S. Zilberstein, Meta-level control of anytime algorithms with online performance prediction, in: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, 2018.
[166] B. Mellers, L. Ungar, J. Baron, J. Ramos, B. Gurcay, K. Fincher, S.E. Scott, D. Moore, P. Atanasov, S.A. Swift, et al., Psychological strategies for winning a geopolitical forecasting tournament, Psychol. Sci. 25 (5) (2014) 1106–1115.
[167] M.J. Salganik, I. Lundberg, A.T. Kindel, C.E. Ahearn, K. Al-Ghoneim, A. Almaatouq, D.M. Altschul, J.E. Brand, N.B. Carnegie, R.J. Compton, et al., Measuring the predictability of life outcomes with a scientific mass collaboration, Proc. Natl. Acad. Sci. 117 (15) (2020) 8398–8403.
[168] B.C. Wintle, E.T. Smith, M. Bush, F. Mody, D.P. Wilkinson, A.M. Hanea, A. Marcoci, H. Fraser, V. Hemming, F.S. Thorn, et al., Predicting and reasoning about replicability using structured groups, R Soc. Open Sci. 10 (6) (2023) 221553.
[169] V.N. Premakumar, M. Vaiana, F. Pop, J. Rosenblatt, D.S. de Lucena, K. Ziman, M.S.A. Graziano, Unexpected benefits of self-modeling in neural systems, (2024) arXiv preprint arXiv:2407.10188.

[170] T. Bricken, A. Templeton, J. Batson, B. Chen, A. Jermyn, T. Conerly, N. Turner, C. Anil, C. Denison, A. Askell, R. Lasenby, Y. Wu, S. Kravec, N. Schiefer, T. Maxwell, N. Joseph, Z. Hatfield-Dodds, A. Tamkin, K. Nguyen, B. McLean, J.E. Burke, T. Hume, S. Carter, T. Henighan, C. Olah, Towards monosemanticity: decomposing language models with dictionary learning, Transf. Circ. Thread (2023). https://transformer-circuits.pub/2023/monosemantic-features/index.html.
[171] T.G.J. Rudner, X. Pan, Y.L. Li, R. Shwartz-Ziv, A.G. Wilson, Fine-tuning with uncertainty-aware priors makes vision and language foundation models more reliable, in: ICML 2024 Workshop on Structured Probabilistic Inference {\&} Generative Modeling, 2025.
[172] C. Ferri, J. Hernández-Orallo, Cautious classifiers, ROCAI 4 (2004) 27–36.