Let's Talk AI with José Hernández-Orallo

José Hernández-Orallo[1] and Barbara Steffen[2]

[1] Universitat Politècnica de València, Valencian Research Institute for Artificial Intelligence, Spain,
[2] jorallo@upv.es
[3] METAFrame Technologies GmbH,
barbara.steffen@metaframe.de

"We have to measure what AI is capable of, and we have to measure our dreams of future AI as well."

The Interviewee - José Hernández-Orallo

My Personal AI Mission:
Understand intelligence, with measurement as the main scientific tool for this.

My Takes on AI

Artificial Intelligence: Intelligence is what solves all solvable problems, following the thirteenth-century philosopher Ramon Llull. Artificial intelligence is just the non-biological kind.

Trust: Meeting expectations, requiring a subject A (in this case a human) to have a good model of subject B (in this case a machine) to know where A expects a valid or invalid outcome when interacting with B. If you cannot anticipate that you cannot have trust.

Explainability: This is easily confused with ex-post justifications, and the area of XAI needs to be crisper in what a true explanation really is. I prefer to talk about models of AI systems that have explanatory and predictable power. Predictability is closely connected with AI evaluation, and predictions are usually easier to check than explanations.

Essential Elements of Human Capabilities: Human capabilities are determined by evolution and culture. We have a very sophisticated perception system, inherited from the primate family, and very advanced social capabilities. Then, of course, language capabilities are a more recent innovation in evolution, which boost the potential for communication, reasoning, culture, etc.

The Interview

Barbara *Today I have the pleasure of interviewing Professor José Hernández-Orallo. Please introduce yourself and your relationship to artificial intelligence.*

José Thank you for having me here today. My relationship with AI dates back quite a while. I've been interested in intelligence since I was a teenager. I remember reading books about anthropology and hominids. At some point, I began to ponder what it would take for a machine to replicate some of these behaviors. That's when I became interested in artificial intelligence. Over the past 20 years, my focus has been more on understanding than developing new AI systems, although I have done a bit of that as well. I am mostly interested in understanding what kind of capabilities these systems can have. This is the goal of the area of AI evaluation, which is a significant topic these days [2, 16, 7]. Especially with general-purpose AI, I'm interested in understanding what these systems are capable of and why they sometimes fail so catastrophically. These are the things that I am currently working on [11].

"If a system can do everything for you, what's the motivation to work hard, learn, and do things yourself?"

Barbara *Is it about the distinction between the intelligence we see in humans and the intelligence of systems?*

José I would go even further back to compare the intelligence in animals, non-human animals, or even children with AI systems. Especially these days, with all these large language models, we often compare these systems with humans. However, in other areas, such as robotics or reinforcement learning agents, it is much more interesting to compare them with a rat or an insect. I believe we gain a lot of insight from these comparisons. But I don't think it's accurate to say that we have systems today that have the intelligence of a rat. I think these comparisons are too simplistic. Instead, there are many tools that have been developed for understanding animal behavior and human behavior. I've been inspired by animal cognition and psychometrics. I think that's where we can find a lot of tools and ideas to evaluate AI systems as well.

Barbara *You wrote a whole book about intelligence. What is your understanding of what intelligence is or how intelligence can be measured? Or is it too complicated to break it down like that?*

José It's very complicated. Intelligence is a term I don't even define in the book. The book primarily highlights all the things that don't work, more than trying to find solutions [5]. Of course, there are proposals and frameworks for evaluation and ideas that show promise. But there's a long debate about what intelligence is in humans, in non-human animals, and how we compare humans in our evolutionary history. When we try to compare humans with machines and what these machines can do, we find that the term intelligence is used very differently for humans and non-human animals. I prefer to use the term cognition

because it's much broader in many ways We don't assume that the system is intelligent [8]. It has some cognitive capabilities and behavior, and we want to understand how the system works. In my book, I try to be comprehensive in terms of understanding and evaluating the intelligence of this diverse range of systems that we can call intelligent. And there are more open questions than answers. But I think that a more holistic approach is required. We need to draw inspiration from the behavioral sciences, from the old disciplines of cybernetics, and so on. I think we need that approach more than many of the approaches that we see in artificial intelligence today, which is basically trying to test the system with a benchmark. That doesn't give you much insight.

Barbara *Are there one or two specific AI-related research questions that you're currently addressing?*

José Yes, there are two questions I've been trying to understand, and of course, they are related to the notion of intelligence. The first one is that, instead of talking about general intelligence, a term introduced in psychometrics by Spearman about a century ago [12], I try to disentangle the notion of generality. What does it mean to have a system that is general compared to a system that is capable? A few years ago, when I was trying to develop metrics of generality people didn't understand. They said, if it is more general, then it is more capable. No, not necessarily. You can have systems that can do a lot of things, but not very well. Not with depth, as a psychometrician or a psychologist might say. A lot of breadth, but not a lot of depth. The system can do a lot of things, but not many sophisticated things. For instance, when you look for an assistant, you want someone or something that helps you with a lot of things, but don't ask that assistant to do very complicated things. Basically, just fetch this for me, or do this for me, or write this up, or something like that. Simple things, but a diversity of things. And this is exactly what we have now with large language models. We have general systems that are not very capable, but can do a lot of things. Some of the things better than others, but in a way, none of them extremely well. Of course, you can find specialized systems that are better than a large language model for each of these things. The interesting thing about a large language model is that the same system can do a lot of things. For the first time, we have these general systems. And this has nothing to do with AGI, with having human-level AI. It's basically that for the first time we have systems that are really general, that we can ask many things to, and they can do many tasks. This generality is out of the box. Trying to introduce metrics of generality is one of the challenges I've been working on [9]. Then the second, related question is that you cannot define generality without a notion of difficulty. Again, this also comes from psychology and psychometrics. You can say, "Oh, I can do additions." Okay, fine, up to how many digits? Because you can't do all additions. Nobody can. This is not even feasible. There are additions with a number of digits that you will never finish. So, what do we really mean? We need a distribution of additions that we expect to happen. That's one approach. But another approach is, "Okay, I don't know what distribution you're going

to experience." Maybe sometimes I get an addition with three digits. Maybe tomorrow I have an addition with 10 digits. I don't know the distribution of additions. I can calculate an aggregate of that, but that's not very meaningful. What is really meaningful is whether I can just order all of these additions in terms of their difficulty. And I can say, for instance, that maybe the mean of the number of digits of the two summands is a good indicator of the difficulty of addition. Of course, there's some carrying going on and some long additions that are, of course, easy, such as 11111, plus 22222. That's a very easy addition. So, it's not that simple, but you can get proxies of difficulty and you can say, "Okay, I can do addition in 10 seconds up to 5 digits." And for an addition of 30 digits I'm not going to get it right in 10 seconds. That's a notion of difficulty that allows me to determine a notion of capability. So, let's say my capability of adding numbers in 10 seconds is 5 digits. Now, independently of the distribution of additions that I'm going to see in the future, you can more or less predict whether I'm going to succeed or not for a future addition. I think that's the kind of capabilities that derive from these difficulty metrics that we need in AI. And again, in psychometrics, with an item response theory you can calculate all of these things, and oppose ability to difficulty. As a result, these two areas the generality of cognitive systems and the notion of difficulty, are closely related, because they lead to proper definitions of capabilities.

Barbara *Do you think that we have an idea of what is difficult for machines? Is the notion of difficulty different for humans and AI? How would you translate between these two?*

José We typically associate the difficulty of a task with the capability of the kind of subject [4]. For instance, making a cup of coffee is easy for a human but hard for a machine, while solving complex integrals may be easy for a machine but difficult for some humans. Difficulty is subjective to a point. But when you fix the task, and look into the instances of the task, you find that despite the different capabilities, there's some instance performance correlations in terms of what you see in general systems [9, 6]. For instance, going back to the example of addition, if you compare a language model with humans, you see very similar patterns of failure when the numbers get larger. And a language model, like a human, finds some instances more difficult than others. In terms of tasks, however, that depends more on the training, making what is difficult for one person easier for another person. In many cases, we can find some kind of common ground. This is tricky, but I don't think it's impossible. Actually, one of the things that I introduced in my book was a kind of universal difficulty scale, which explains why this is subjective to a point. When you build a system that is general, then you find some commonalities in terms of difficulty. That's something that I hope that at least experimentally we can start seeing in some of the new general systems that we're building in AI [3].

"Of course, we need to be careful about what we build, as we're essentially creating new beings."

Barbara *In your opinion, what role does trust play in the adoption of AI?*

José Trust plays two roles, both positive and negative. Of course, if we don't trust AI, we will use it less than we should. There are significant safety issues with some technologies, and people react against these technologies. Even public opinion can turn against some products. That's why companies are careful about what they do. Maybe not careful enough, but at least they care a little bit that these systems do not do things that people could criticize. That's one thing. The other thing is when they create too much trust. And that's more dangerous. I believe it's better for people to distrust systems rather than overtrust them. Ideally, we should know exactly where a system is reliable and where it isn't. However, this is often difficult to determine, particularly with large language models. For instance, when we pose a question to ChatGPT, we can't predict the response [13]. Not only is the outcome uncertain, but we are often unsure of its accuracy. Sometimes, we might ask the system to write something for us and be pleasantly surprised by the result. But if we start relying on machines for a lot of things, and we think that they can do them well, and at some point they surprise us by doing something wrong, or something really, really wrong, then this is a big problem. So, we need to calibrate trust. Of course, it will be ideal if these systems were consistent. But they are not. Sometimes you ask the same thing or some variation of the same question, and you get something that is rubbish, basically. So, this creates a problem of expectations [15]. The users of these technologies don't know what to expect. And the learning curves are quite long. When you start using ChatGPT, the first thing is, oh, look, I asked it to prove this conjecture as a poem. And I get it. Oh, this system must be fantastic. And then it fails with additions of five numbers. You're really disappointed because you don't expect a system that is able to write this prose and all these poems, even solve some easy differential equations, and then fail on a simple addition, when a calculator, another machine, does this perfectly. All of this breaks our schemas about what to expect from a machine. And that creates a problem of trust. But sometimes there's over-reliance on the system. You think: Oh, that's so cool. They can do so many things. I can just write a summary. I can send e-mails using these tools. And then I find out I screwed it up because I relied on this system. So, this is a major problem at this moment. But if we had to choose, I would choose to have less trust than they really deserve.

"The existential risk isn't so much about these systems getting out of control, but about human disagreement on important questions about our future. Are we going to remain as we are? Are we going to create other systems? Are we going to grant them rights?"

Barbara *So, it's essentially about educating users to remain a bit skeptical and to better understand and reflect on what the system can actually do. Where can it complement us? And where do we currently have a distorted idea of what it can do? Given your mathematical example, it seems strange to us that ChatGPT*

can write these wonderful poems, but it can't do calculations that we learn in elementary school. So, we have the misconception that what is difficult or easy for humans must also be difficult or easy for the system.

José Yeah. The problem is that we cannot give all this responsibility to the user. This happens with computers and with any technology. We try to adapt to the technology. And, okay, the first time you use a computer, you say, well, what can I do with this? You download a new app and you try to adapt. And even if you have a digital assistant that is not fueled by a lot of AI, you know, okay, if I ask a question "play this song for me", this is going to work. But if I ask some other questions, such as the meaning of life, you're going to have some kind of prerecorded or prewritten answer for these kinds of questions. And you have to learn all of that. And then, you start to know when these digital assistants are useful or not. But this is a lot of effort, a long learning curve. With a system that has been designed to be an assistant, that's still okay. But a system such as ChatGPT has been adapted to be an assistant from a raw language model. Things are much more complicated. And we cannot ask humans just to build a perfect model of what ChatGPT can do.

So that's why one of the things that I'm doing research on is how we can build this kind of external model (an assessor) of what the system can do and can't do [10]. And then use this in a kind of a monitor or kind of an advisor telling you, okay, the system is going to fail at this or not. Because there are questions about what people call scalable oversight. So how can humans know whether the system is correct or not? And with these more powerful systems, it's becoming more and more difficult to know, even to ask an expert, is the system correct or not [15]? In many cases, it is even debatable what the ground truth is, especially about things that are a little bit more vague about society or things like that. Even scientific facts, you can argue with some of these systems about what you get. That's why I think we need more assistance. We can't rely on the regular user to build a perfect model of when the system is correct or not. Because that would entail that you know, at least in many cases, more than the system knows. And then the system wouldn't be very useful. Instead, we want these systems to be very good generators of things we cannot easily do. People call them generative AI. But apart from generators we also need verifiers [14]. That's something where some other areas of AI and computer science are much better than the current trend of transformers and generative AI.

Barbara *Do you have any specific measures in mind that could help ensure ethical use of AI?*

José Well, there are many things. There are so many problems about the ethics of AI. One thing is the way in which these tools are ethical in the first place. And of course, when I mean ethical, it's not that we could talk about moral machines or something like that. Basically the use of these systems could lead to discrimination or inequalities or even increasing the inequalities that we already have. All of these issues are now on the table. I'm happy that there are many discussions today about AI around this. But there are also many things as well

that go beyond just whether a system is politically correct. It's more about the geopolitics. I think that if we want to deploy AI in an ethical way, we need a more inclusive AI community. At the moment, it is not very inclusive. Not only is it dominated by a few countries, but a few areas of a few countries, with some big tech having an oligopoly on this at the moment. Some profiles of gender and race have dominated the discourse. This has to be changed from the inside. Also, the data that we use to build these AI systems is completely biased because humans are biased. Sometimes this bias is amplified. However, having said all of this, I think that we have a big opportunity. I see some human judges making some kind of resolutions, and I see a lot of biases in them. It is quite rare that you analyze this in an evidence-based way. We humans discriminate. We do that all the time. We are biased. And society is completely unfair in many cases. But we don't often use data to analyze that. For the first time, when we feed all this data to machines and build a model, we really see how unfair society is, all these biases that people have. We basically reidentify the biases in the machine. We can measure them. And then we try to correct them. But the correction is very difficult when the data is biased in the first place. You have trained the system on a lot of rubbish that you have just gathered from the Internet, not especially the best Internet sources you can find. So, what do you expect? Basically, you're going to replicate all these biases. But in a way, I would see it the other way around, like a mirror allowing us to identify and look at all the biases that we have in society. And this is a way in which we can just point out that this is happening in our society. AI is basically resurfacing all of these biases. Today, we try to apply all these new laws for AI systems. Okay, but I say, let's apply all the old laws for humans as well. Especially when they are not fair in their decisions: politicians, judges, police. I think that this should be applied to AI and to humans. Perhaps not in the same scale because AI has a power of replication, possibly having much more effect than a single person. I understand that people are concerned that if a system is biased, that can have more effect than if a single person is biased. But in the end, we have to solve the problem of people being biased in the first place.

Barbara *Okay, so the biases of AI are more systematic, but at the same time AI systems also reveal the existing biases of society and make them transparent, because all the biases of the AI are the result of the "real world" data that is fed into the large language model.*

José Yeah, I think that it is a mirror of society. And having a mirror to see yourself, I think it's very, very insightful. It can highlight the real problems and its sources. When you are trying to select the "good" sources, you realize that this is really complicated. The problem of bias and ethics requires people who are experts in ethics. Sometimes people in AI who are really mathematicians or physicists don't know anything about ethics, as a discipline. At least the engineers, especially computer engineers, they usually have a course in the ethics of their profession. And they know the user is very important. The user is a human, so engineers have to build systems that are basically serving the purpose

of humans. But in many cases, some of these big tech companies have a lot of engineers that were not trained as engineers, they're just mathematicians and physicists. They have been trained on formulas, but not on people. They don't know how to act with people. So, there's a lot of things to improve there, in terms of the people themselves. That was related to the start of the answer to the previous question; we have to change AI from the inside.

Barbara *Now looking into the future and especially the possible future capabilities of artificial intelligence on a scale of 1 to 10, where 1 describes the artificial intelligence tools like ChatGPT that we know today. And 10 refers to artificial general intelligence that surpasses human capabilities. What do you think will be possible?*

José With no timeline, I think that anything that is computable is possible. I think the only limits are given by physics. That's what I see. And of course, humans are quite limited in many ways. So, it's just a question of time but it's also a question of what we want to do. It is not a given thing that they're going to build some kind of system that is much more powerful than humans. That's something that we have to decide. And we have to decide what kind of system we want to build. First, because it might be dangerous. And second, because it might be unethical. In biology, we agreed we are not going to play with DNA and do this kind of chimeras mixing a cat and a dog: "Oh look, how cool, is this new animal we have created." Because maybe this animal starts suffering. You can use an elephant and a mammoth DNA and then recover more mammoth DNA and try to see if, in a couple of generations, you have a real mammoth. These things are basically unethical. But creating something that goes beyond us, that's kind of a dream. But we have to be very careful about our dreams. So, whether we want to reach that 10 in a scale of 1 to 10 is a decision we have to make. And there are different choices. It's not just a single scale. We have to be careful in choosing from the infinitely many options that are more powerful than us.

"We're not investing enough effort into understanding a potential cognitive atrophy, similar to how we've physically atrophied due to over-reliance on technology like cars."

Barbara *Going back to the beginning of the interview where we talked about intelligence. Do you think we have an idea or a common understanding of what it is and when we have reached it? Or are these still very subjective concepts and ideas at the moment?*

José There's no consensus. There's no science at the level of understanding intelligence. We are playing with something that we don't understand well. And now it's quite trendy again to talk about nuclear physics and the first Manhattan Project. But how much did they know about what they were building compared to how much we know about what we are building today? I think that we are far worse with AI than with nuclear physics. And nuclear physics sounds very scary.

But AI might also be scary. Not that much in terms of creating something that gets out of a lab, but creating something that's going to have a lot of implications for humans, starting with human cognition. We don't understand things well. And we are trying to play, "Okay, let's build the next generation of this system. Let's see what happens. Oh, cool. Oh, no, it's not that cool." That's the way we are today. And we have these fancy scaling laws. We just scale the number of parameters or FLOPS and get these new capabilities. Is that the only thing that we know about intelligence? That we just scale the size or computeof the neural network and we get more capabilities? Is that all that we know about intelligence these days? If that's all we know, I think that we are at a really, really basic level of understanding to try to popularize and develop a technology for which we don't have the science.

Barbara *Looking into the future, how will these developments continue? Where would you place yourself on the spectrum from dystopia to utopia?*

José I'm an optimist. I think there are more positive things than negative things in AI. But I put a lot of emphasis on existential risks. While I don't necessarily believe these risks are highly probable, my concern stems primarily from our inability to accurately gauge their likelihood. When faced with significant dangers whose probabilities are difficult to estimate, even if we believe they are low, it's crucial to invest more effort into understanding them. Scientists have been working in the past decades to estimate the probability that an asteroid would destroy life on Earth. We now have good estimates of how likely this is because we have seen this in the past. We know that a big one happens every 100 million years. In AI, this is something that we need to calibrate too. Particularly, we need to estimate the probability of these significant risks, which we currently can't do accurately. Because we're developing technology without a solid scientific foundation, which is concerning. Focusing on existential risks doesn't mean there's a divide between practical ethical concerns about AI today and future AI problems. These two aspects are interconnected. Paying attention to these major issues requires a better understanding of AI, which also aids ethical considerations and AI usage. In this continuum of issues, my primary concern is how AI will alter human cognition. We're not investing enough effort into understanding a potential cognitive atrophy, similar to how we've physically atrophied due to over-reliance on technology like cars. This, to me, is a major concern, perhaps even more so than some of the other issues people are discussing.

"I believe it's better for people to distrust systems rather than overtrust them."

Barbara *Is there already research on this? Or is this something that is often overlooked?*

José More people are discussing this, especially with platforms like ChatGPT being used by millions, including children [1]. This could significantly affect not only their cognitive development and problem-solving abilities in the future,

but also their motivation. If a system can do everything for you, what's the motivation to work hard, learn, and do things yourself? We've seen similar effects in the physical world and with social media. This could escalate rapidly.

Barbara *Could this lead to a decline in our intelligence?*

José Indeed, a significant part of intelligence is innate. But if you don't use it, especially if you don't see a motivation for using it, it could atrophy. We've become so reliant on technology that we would be helpless in a natural environment. We need to use these tools to empower ourselves, even if it means some of our abilities might atrophy. However, there might be situations where this goes too far, especially if some day in the future we no longer need to work. Because of this we need to return to the Enlightenment principle of understanding the world for its own sake, not just for professional training. But this requires motivation, which could be challenging in a world where work is no longer necessary. The message needs to change, and that's a challenge.

Barbara *Is there a specific area of research you would like to see addressed more from a multidisciplinary perspective?*

José Yes, especially the impact of AI on cognition. Psychologists and behavioral scientists are starting to incorporate AI into their research, which is beneficial. They understand cognition, particularly human cognition, and how it can affect mental health, education, and the workplace. Economists also play a crucial role. This needs to be a collective effort, as AI is the technology of the century. It's going to change everything.

Barbara *What is your vision for AI?*

José My primary goal, as I've mentioned since the beginning of the interview, is to understand intelligence. AI is the main tool we have for this. I hope it will give us more insight into what intelligence is and the different types of intelligence that can exist. Evolution has given us some types of intelligence, but there might be others that we haven't yet discovered. It's fascinating to think about all the different kinds of intelligence we could create. Of course, we need to be careful about what we build, as we're essentially creating new beings. But from a scientific perspective, it's incredibly exciting.

Barbara *As we move forward, we might encounter unknown unknowns. So, there may be an intelligence out there, now or at some point in the future, that we as humans are not capable of recognizing. Would we notice?*

José We will have to make some significant decisions. We will have to decide whether to preserve Homo sapiens for millions of years on Earth as a reserve or whether the species transitions into something different. There will be reactions to these changes, and there will be disagreements, geopolitical problems, and more. But we will have to navigate these challenges and see what happens. The existential risk isn't so much about these systems getting out of control, but about human disagreement on important questions about our future. Are we going to remain as we are? Are we going to create other systems? Are we going

to grant them rights? These debates are already happening and will become more relevant in the years to come. Politicians aren't discussing this yet, but they will.

Barbara *Is there anything else you would like to add?*

José No, I'm just an optimist. I believe this is one of the most exciting times for science, and I feel privileged to work in AI. However, this excitement shouldn't lead us to rush. We're close to realizing the dream that early AI pioneers had decades ago, but we need to proceed with caution and focus more on science and less on technology.

Barbara *And on the societal impact, right?*

José Yes, of course.

Barbara *Thank you very much for your time, José, and especially for your perspective on intelligence. I am excited to see what will happen in the next few years. And I look forward to the progress towards a better understanding of AI and our responsibility to consciously steer the next steps in the desired direction. Have a great day!*

José Thank you very much.

References

1. Bai, L., Liu, X., & Su, J. (2023). ChatGPT: The cognitive effects on learning and memory. Brain-X.
2. Ryan Burnell et al. Rethink reporting of evaluation results in AI. Science380,136-138(2023).DOI:10.1126/science.adf6369
3. Ryan Burnell, Han Hao, Andrew R. A. Conway, José Hernández-Orallo: Revealing the structure of language model capabilities. CoRR abs/2306.10062 (2023)
4. Desender, K., Van Opstal, F. & Van den Bussche, E. Subjective experience of difficulty depends on multiple cues. Sci Rep 7, 44222 (2017). `https://doi.org/10.1038/srep44222`
5. José Hernández-Orallo: The Measure of All Minds: Evaluating Natural and Artificial Intelligence. Cambridge University Press 2017, ISBN 9781316594179
6. Hernández-Orallo, J. Unbridled mental power. Nature Phys 15, 106 (2019). `https://doi.org/10.1038/s41567-018-0388-1`
7. José Hernández-Orallo, Marco Baroni, Jordi Bieger, Nader Chmait, David L. Dowe, Katja Hofmann, Fernando Martínez-Plumed, Claes Strannegård, Kristinn R. Thórisson: A New AI Evaluation Cosmos: Ready to Play the Game? AI Mag. 38(3): 66-69 (2017) José Hernández-Orallo: Evaluation in artificial intelligence: from task-oriented to ability-oriented measurement. Artif. Intell. Rev. 48(3): 397-447 (2017)
8. José Hernández-Orallo, David L. Dowe, M. Victoria Hernández-Lloreda: Universal psychometrics: Measuring cognitive abilities in the machine kingdom. Cogn. Syst. Res. 27: 50-74 (2014)
9. Hernández-Orallo, J., Loe, B.S., Cheke, L. et al. General intelligence disentangled via a generality metric for natural and artificial intelligence. Sci Rep 11, 22822 (2021). `https://doi.org/10.1038/s41598-021-01997-7`

10. José Hernández-Orallo, Wout Schellaert, Fernando Martínez-Plumed: Training on the Test Set: Mapping the System-Problem Space in AI. AAAI 2022: 12256-12261
11. Wout Schellaert, Fernando Martínez-Plumed, Karina Vold, John Burden, Pablo A. M. Casares, Bao Sheng Loe, Roi Reichart, Seán Ó hÉigeartaigh, Anna Korhonen, José Hernández-Orallo: Your Prompt is My Command: On Assessing the Human-Centred Generality of Multimodal Models. J. Artif. Intell. Res. 77: 377-394 (2023) Xiting Wang, Liming Jiang, José Hernández-Orallo, Luning Sun, David Stillwell, Fang Luo, Xing Xie: Evaluating General-Purpose AI with Psychometrics. CoRR abs/2310.16379 (2023)
12. Spearmen, C. (1904). General intelligence objectively determined and measured. American Journal of Psychology, 15, 107-197.
13. Lexin Zhou, Pablo Antonio Moreno Casares, Fernando Martínez-Plumed, John Burden, Ryan Burnell, Lucy Cheke, Cèsar Ferri, Alexandru Marcoci, Behzad Mehrbakhsh, Yael Moros-Daval, Seán Ó hÉigeartaigh, Danaja Rutar, Wout Schellaert, Konstantinos Voudouris, José Hernández-Orallo: Predictable Artificial Intelligence. CoRR abs/2310.06167 (2023)
14. Yunxiang Zhang, Muhammad Khalifa, Lajanugen Logeswaran, Jaekyeom Kim, Moontae Lee, Honglak Lee, Lu Wang: Small Language Models Need Strong Verifiers to Self-Correct Reasoning. CoRR abs/2404.17140 (2024)
15. Zhou, L., Schellaert, W., Martinez-Plumed, F., Moros-Daval, Y., Ferri, C., Hernandez-Orallo, J. "Larger and More Instructable Language Models Turned Less Reliable", Nature, 2024, to appear.
16. `https://aievaluation.substack.com/`