TOMER ULLMAN: So yeah, I'm going to spend the rest of the tutorial talking about Amazon Mechanical Turk, although some of the stuff just applies to, in general, sort of making large scale experiments on people and crowds and things like that. There are other alternatives. I think Google has some in-house things. There's like crowd flower or crowd clicker or things like that. But most of the psychologists I know that are doing stuff online and large scale things use Amazon Mechanical Turk. So I'll be talking about that.

This is a crowdsourcing platform which is designed to do all sorts of small tasks, not necessarily psychophysics. It was invented in around or built in around 2005. And the signature tagline for it is "Artificial artificial intelligence." So these are all sorts of tasks that you wish computers could do. You don't want to put people through it but computers can't do it yet. So let's have people do it for now. Amazon sort of invented it for in-house purposes to get rid of duplicate pages.

They really wanted an algorithm to do that for duplicate postings, but there really wasn't one. So they just paid some people very little money to do a lot of these pages. And then they figured, wait a minute, a lot of companies probably want this sort of service so they offered it to the general public. Do people know what the original turk was, the original Mechanical Turk? Who doesn't know, raise their hands. OK, so Mechanical Turk, the name comes from this 18th century mechanical contraption called the turk, which was a chess playing device that supposedly ran on clockwork and could beat some of the finest minds in Europe.

I think at some point it played like the Austrian duchess or whatever it was the empress and Napoleon and things like that. And it floored people, how could this work? And you know, the inventor said, well, I've invented a thinking device. And the clockwork just solves it. And a lot of people of course figured that it must be a hoax but they couldn't be sure exactly how it worked. And it was, of course, a hoax. All of these gears and boxes and clockwork is just designed to distract you from the fact that you could fit a person inside. I mean, people have thought of that, obviously, but it was cleverly designed so that you couldn't quite find it.

And nobody knows for sure because the original work was destroyed before people put forward the exact hypothesis, but the consensus now is that it must have been just some person inside making the moves. At some point people have suggested it must have been a well-trained child or a small person or something like that because the compartment must be really small. That's not thinking right now. It was just some more magic, but not magic in the Harry Potter sense magic, in the stage magic sense.

So, what sorts of tasks do people run on Amazon Mechanical Turk? Well actually, the majority is not psychophysics or psychologists. There are a lot of companies using Amazon Mechanical Turk to do things like, hey, garner positive reviews for us. Go to this website and write something nice about our product. Or you know, translate some text for me instead of hiring some professional to do it, most people who-- there's a lot of people on Amazon Mechanical Turk that can probably translate something for you from English to Spanish and back, English to Chinese and back, things like that. And they can do it for much less money than you would need to pay a professional translator.

Another thing that may be more up your alley is, you know, you've heard a lot about supervised training and things like that, and big data sets that are used to train things like convolutional neural networks. The supervised part usually comes from somewhere. Somebody has to tag those images. Somebody has to go over a million images and say dog, not dog, dog, not dog, not dog, two dogs, one dog. You want somebody to go ahead and do that. And you don't have artificial intelligence to do it for you, yet. I mean, CNN's are getting better but someone had to tag it for them.

And that's the sort of thing that you would use Amazon Mechanical Turk for. Look at this image. Tell me, is there a social interaction or isn't there? Now do that for the next 100 images. Let's get 1,000 people to do that. You get the sense of like why you would want to crowdsource this kind of problem. But there's also, you know, just psychology, psychophysics, the sort of thing that you would bring people into the lab to do but you don't want to do for various reasons.

So the idea is you could collect a lot of people responding to stuff on the screen right, exactly the sort of thing that you would bring them into the lab and measure something that they're doing to a screen. You could just have them doing that to the screen at home, or wherever the hell it is that they're doing this thing. So let's see, you could do things like perception. You

know, just look at this. Is it a dog? Yes or no. Or rather things like the Stroop task. If nobody had invented the Stroop task, you could do that on Amazon Mechanical Turk and get famous. I was going to say rich and famous, but just famous.

You could do various attention tasks. You could do things like learning and categorization and bias. Learning things like, here's a tufa, here's a tufa, here's a tufa. Is this a tufa? I mean, it's very easy to do that on a screen. You can do that on Amazon Mechanical Turk. You could do things like implicit bias. A lot of social psychologists are interested in that sort of thing. Let's see, what else. You could do things like morality, the trolley problem. You could do things like decision making, and economics, and prisoner dilemmas, and making predictions, and tell us which one of these two movies based on trailers do you think will win the best actor award and things like that.

How would you actually run something on Amazon Mechanical Turk? The first thing you would do is register as a requester. You would go to requester.amazon.-- I'll send you the slides. You can just click on that link and it'll take you to the page to register as a requester. You would then you do one of two things. You could either do the vanilla version which is to just use the Amazon template. Amazon has made it very simple for you as a requester to bring up a new experiment and just say, I want to ask a simple set of questions. Now go. And there are some parameters that you need to set and I'll show you those in a second, what they do.

And if you're only interested in very simple things, like fill out this box or click on one of these two images or something like that, that's perfect and it's fine. And Amazon takes care of a lot of things for you like saving the data. You don't have to mess around with like SQL databases and things like that on your own. The other thing that you could do is to point people to an external website. Then you would have to host it somehow. The advantage there is that then you could do any sort of fanciness that you want. You could show them custom animations, have them play a game, record how they're playing that game, send them new things based on how they're playing that game.

Or you could do things like, you wait until you record two people. You record two people. If you just recruit one person, it just says waiting, waiting, waiting. Now you wait for the next person. Now you have them pitted against one another in some sort of game. This has become more popular in economics. That's not the sort of thing you could do with the Amazon template. But if you're good at coding or you can hire someone that's reasonable at coding, you can do that by pointing them to an external website. And we can show some examples of that.

Then once you decide-- you register as a requester. You decide which one of these things you want. You build your website, either you build the external website or you just use the Amazon template. And then you test it on a sandbox. Don't run your experiments live immediately. I'll be giving sort of tips throughout the thing. This might be redundant for some of you but not redundant for others. So there's a sandbox where you can just run it and get sort of false responses that don't count exactly where people can sort of fill it in. You might want to used that before you go live with 1,000 people and say, oh god, I miscoded that variable and nothing is working. So test it ahead of time.

And then, once you're finally, finally done, you would submit it. You would just click, you know, submit this thing. You would pay the money to Amazon. And you also want to announce it on several Mechanical Turk forums, which I'll get to in the end. These are very helpful people. They're very nice to you if you're nice to them. And it gets you a lot more results very fast. OK, let's see. Why don't I show you an example-- let me show you an example of what a requester page looks like, just to get a sense of it for those of you who have not seen this sort of thing before.

You can see our experiment is almost finished. I asked for 100 people on that. So let's go to something like create. So this is what the requester page looks like. Here's all sorts of projects that I run on Amazon Mechanical Turk. And you would do something like new or you would copy something from something old that you already did. Let's just edit that and show you some examples of what you can do. You give a name, you know, a title for your own internal title, like AI estimate. You then give a title that Amazon Mechanical Turkers see, something like artificial intelligence estimate, short psychology study.

There, you probably want to give a time estimate. Turkers would prefer it if you give them a time estimate. They care about their time. They care about their money. They do like doing psychology. They don't like filling in endless bubbles, you know, the standard psychology things where you rate 100 things like, I feel this way or that way. Don't do that. But the sort of fun psychology they're actually on board with. It's much more fun than writing show reviews.

You might want to give it a nice title that will entice them, an honest description, like you know, you'll answer a few simple questions on you'll watch a movie and then answer two questions or things like that. Key words like easy, fun, something descriptive about the task, like AI, short. Again, this is sort of luring, and it's good to do that if you're honest about it. Some

people do things like easy, fun, 100 pages of filling in bubbles and things like that. Don't do that. They publish it in the forums, like, this is a lie. Don't do that. This is where you say how much you want to pay per assignment and we'll get to how much you should pay per assignment. You'll notice it's very little. You're paying these people very little.

How much assignments you want per hit. Hit is just the name for your task. How much time you are allotting them per assignment. So someone has accepted your hit, now how long do they have to carry it out. You might think, oh, my task only takes two to three minutes, so only give people two to three minutes. I don't want them like taking the hit and then going and drinking some coffee and then coming back to it or something like that. Consider still giving them a whole bunch of time, because a lot of the time, you'll find that people if they see that there's only a five minute mark on it, and if they don't completed in five minutes they're sort of concerned, like, what is this thing? And if I won't get done through it in five minutes, you know, I'll get disqualified or something like that.

Give them some time. Give them more than the ample time to finish this thing. You can, yourself, keep around some timer within the external website or something like that, or they actually actively participating right now. How long will this hit stay up for? And auto approve, and things like that. I'll talk a little bit about rewards and incentives and things like that. Obviously, they care a lot about things like money. They care a lot about things like how much you're going to pay them. They care about doing it quickly so that they can move on and get more money.

They care about it being somewhat fun, but that's not such a big deal for them. And they care about getting it approved quickly. OK, so you as a requester, you will get reviewed on various forums and things like that. If you get bad reviews, people don't want to do your things. One of the things that people care a lot about is something like getting approved quickly. And quickly can be in a day or two, something like that. They don't want to have to wait two weeks for that $0.50 that you were supposed to give them. You can do that if you want to. You have the power as the requester.

But if you want to incentivize people, try to make sure that you approve them quickly and let them know that you're going to approve them quickly. So that was just a very general statement. So who are the sort of people that are on Amazon Mechanical Turk? Have people read these sort of papers and things like that? Do you know more or less? Some of you do. Some of you don't. The use in India make up about 80%-- by the way, this is in flux. Like a

study came out two years ago about this sort of thing. It's changed since then.

But the study two years ago, the US and India make up about 80%, with the US taking up 50-something percent, India taking up the rest. There are slightly more females than males on Amazon Mechanical Turk, and at least the US population is biased towards young and educated people, educated meaning a bachelor's degree. It's certainly more representative of the general population than just hiring, you know, college students during their bachelor's degree. But it is still skewed, keep that in mind.

It's skewed towards, basically, the sort of population that you would expect to find on the internet in the United States. OK, so somewhat younger people, somewhat more educated. In general, you might want to look at something like Mason and Siddharth were looking at these sort of things. I can send you the links later. I was talking a little bit about payments.

There's been a whole load of studies looking at querying the Amazon Mechanical Turk pool. Who are you? What's your education? Why are you doing this? And then they ask why are you doing this in various ways. Are you doing it for fun? Are you doing it to kill time? Are you doing this as a supplementary thing? They're in it for money. It's very obvious that they're in it for money. That's OK.

And keep that in mind when you post hits. You have a lot of power as a requester. You have a ton of power. You have a ton of power to dictate the terms of what you're going to pay them. You have a ton of power to reject their work. Once they basically do the hit for you, you can then go back and say, you failed this question or you didn't quite get what we wanted, or you actually did the study two months ago but I didn't implement checks for that, I just asked you, did you take the study before and they didn't remember. Something like that.

And then you reject their work. And if you reject it too many times, then they get banned. First of all, if you reject, they don't get paid. If you reject it too many times, they get banned. These are people that are doing it either as supplementary or as their main income. I don't know--this may seem obvious to some of you. If this doesn't seem obvious to at least one of you, then I'll count it as worthwhile to stress this. You don't care about the $0.20. These people do care about the $0.20.

And again, it's not because they're necessarily from poor economical backgrounds, but they're in it for the money and that's what they're doing this for. Try to give them fair pay. And we'll stress that again and tell you what I mean by fair pay. Try not to reject them. OK, except in

extreme situations. Even if they failed. Even if they didn't do your catch question. Even if you think that they just zoomed through this or something like that, that's usually on you to catch that, as a psychology researcher. You're not a company. You're a psychology researcher. Try not to reject people.

Make sure you have some ways set up ahead of time, and I'll get to that, to know who to reject. But don't actually reject people except in really extreme situations. Something about payments is that Amazon takes up about 10. They actually raise this to 20% to 40% of the payments. And this is what I was going to say. There have been some attempts within the psychologists that has been doing Amazon Mechanical Turk studies, and this has also been fueled by the community of Amazon Mechanical Turkers, or Turkers, to establish some sort of guidelines for minimum pay.

So if people come into the lab, there are some guidelines on what you're supposed to pay them. There are no exact guidelines. There's no enforcing guidelines, at least not-- maybe there are within particular universities but there's no cross university one guideline to tell you you have to pay people this much. And that might be tempting to say, oh, I'll just pay people, you know, the minimum I can get away with. I mean, if I can pay people $0.05 to do a 10 minute task, I'll do that. Fine, I can get the 20 subjects I need. It's a free market.

We're not trying to live here in some sort of capitalist fantasy of some sort. I'm not going to get into economics too much because I think you guys know that better than I do you don't need my lecturing in that sense. But a lot of people who have looked into the ethics of this recommend that you try to estimate ahead of time through a pilot how long is this test going to take. Based on that, pay them such that it matches minimum wage. Minimum wage being somewhat flux, like, you know, I forget if it's like $10 an hour or something like that. It's probably less than that. But something like that.

I'm not going to tell you exactly how much you should pay them. But try to figure out more or less minimum wage, more or less how long your task takes and pay them according to that. Some general advantages of Mechanical Turk, in case you guys have not been persuaded yet, let me say. You can run large scale experiments. I started running this experiment on 100 people that would have taken me a long time. It's a silly experiment as Nori pointed out. It's not even exactly an experiment.

But I wanted to check how people's responses compared to people in CBMM. First of all, I

wouldn't do that in the lab. So there's that. But even if I were to do it in a lab, getting 100 participants would take a long, long time. And for your more serious experiments, getting 100 participants would take a long, long time. Each one has to come in and you have to talk to them. And you have to explain to them exactly what's going on. And you usually can't run them in parallel, or at least you can run only one or two in parallel. Here we run 100 subjects in an hour.

And that's still amazing. That's still flooring me. And we can do it, so we can do it very quickly. A lot of people very quickly. And what you can do with large scale experiments is usually test some very fine grain things of your model. If your model has some things like, well, I need to show people all of these different things and make all of these different predictions. And I just need 300 people to do that. Or for example, what my dad was presenting yesterday, these minimal images. Did he mention how many people they had to recruit on Amazon Mechanical Turk to do those minimal images? Yes?

It's like it's thousands. I think it's over 10,000 at this point, or something like that. And the reason is because once you've seen that thing, once you've seen the minimal image, you already know what it is. You're biased. You know it's a horse even though a naive participant wouldn't know it's a horse. So you want to make sure that you want 10,000 participants for this thing. You're not going to get 10,000 participants in the lab. No way.

So another thing is that, as I said, even if you're paying people minimum wage and things like that, it's still pretty cheap to get 100 subjects. It's cheapish. The ish is because you should pay people some minimum wage. It's replicable, ish. What I mean here by replicable is not what you might think, which I'll get to in another slide. It's just if you want to hand it off to another person. Someone says, I don't quite understand your protocol, or I don't quite believe it, or I want to tweak it in some way. It's much harder, usually, with lab protocols and things like that. We certainly know that in baby experiments.

Wouldn't it be nice if we could just port, you know, I won't say who because it doesn't really matter, but some experiments of people. They describe their methods in the paper. It's not really that great. Wouldn't it be great if we could just copy paste their experiment and run it with some tweaking. With this sort of thing, you can. I mean, you need to be somewhat on good terms with the person you're asking for, but they can just tell you, oh yeah, sure. Here's the code for my website. Just run it again. Run it with your tweaks and things like that.

There's an ish there and I'll get to it in a second. The participant pool, as I said, it's more diverse. So I was I was harping before about this point that the pool doesn't quite represent the US in general, it's more like the US population on the internet. That's still a lot more diverse than recruiting college students. It's a lot more diverse, let's see, I have here in response to gender, socioeconomic status, geographic region, and age. On all these things that people have tested on Amazon Mechanical Turk, the sample is a lot more diverse, is a lot more representative of the general population, is a lot less weird.

Weird being like Western, educated, industrialized, rich, democracies, which is usually the pool that's been studied in psychology. These pools are a lot more diverse. They're not diverse enough for certain things in social psychology, and that's this paper by Weinberg, where he says, you know, sometimes social psychologists really, really, really want to control for making sure that age is not a factor, or something like that. Or they really want to get it what the population is, or age they think plays a factor, or something like that.

So you need a population where they call it sort of like knowledge experts. You build some pool. You build some pool that you say, OK, the reason this pool exists is because it's representative of the population. And now we're going to go to this pool and just try them again and again and again on many different experiments. It's sort of like your, you know, not exactly private but shared between some universities pool of social psychology participants. They've tested that and they've shown that mechanical turkers are better than those pools in terms of things like attention, filling out the correct responses, and things like that.

So yay Mechanical Turk. But there are some things like implicit biases and things that social psychologists care about, where you don't know if the effect is something like age, or something like that. I don't know if this matters to a lot of you but it's important to keep in mind for those of you who do. Here's this point about will it replicate. You know, some people who are being introduced to Amazon Mechanical Turk, or thinking about it, usually say, yeah, that's fine. But how do I know that people are actually doing what will happen in the lab?

I might do it on Mechanical Turk, but then if I do it in the lab, it won't replicate or things like that. So people have tried a bunch of the psychophysics that Leyla was talking about before, and more. They tried stroop, switching, Flanker, Simon, Posner, cueing, intentional, blink, subliminal priming, and category learning. And they've done a whole lot more. This is just from one study by Crump et al., where one of the et al. is Todd Gureckis, who we'll get to in a second.

And what they find was basically replication on all of these classic psychology stuff, the sort of effects you would expect. The sort of effect sizes you would expect. The only thing that was a bit different was in category learning where you show people something like, this is a tufa. This is a tufa. This is a tufa. Is this a tufa? Where there are different types of learning, type one being easier, type two being of a little harder, type three being much harder. Where the classic finding was something like a graded thing. And for Amazon Mechanical Turk, it was more-- it was really hard for them beyond type one.

Now is that a failure of replication or is that because the original study was done on college students who are young and educated? And this is actually more representative of the population, but this is harder to learn. I'm not quite sure. The takeaway here is that, yeah, it seems like, in general, it will replicate, at least certainly for simple perceptual stuff. Concerns of running things on Amazon Mechanical Turk. And I can send a whole bunch of recent papers that are very nice about it. One of them was specifically-- there's been a bunch of like New York Times papers on Amazon Mechanical Turk in general.

There's been a very recent one a few months ago on using it for psychophysics experiments in particular. It's called the Internet's Hidden Science Factory. It's a very nice paper to check out. And they make all these points about the sort of things that you probably thought about as a researcher but it bears thinking about again, which is, people don't necessarily pay that much attention to your task. You have no control. They're not in the lab.

They give some quotes there, which is, you know, Nancy's employees don't know-- yeah, I think it's Nancy. I changed the name. Nancy's employers don't know that Nancy works while negotiating her toddlers milk bottles and giving him hugs. They don't know that she's seen studies similar to theirs, maybe hundreds, possibly thousands of times. So that brings us, actually, to another thing, which is repeated exposure. This is a big concern. By the way, sorry, I'm going sort of back and forth here because before I leave attention, I want to mention just one thing, which is attention is a problem you want to put in attention cues. And I'll talk about how to do that in a second.

But we've had this a lot with people in the lab. I'm sure that some of you have experienced this as well. You put them in a room because they need to have privacy while they're doing the task, you go in to check on them and they're on their phone. This happens a lot. So attention is something that you want to check in the lab as well. A lot of these concerns are not just

about Mechanical Turk, but it's certainly easier for people in Mechanical Turk to not quite pay attention.

Repeated exposure is a huge problem. And it's a problem for two different reasons. One is that it destroys intuition. I was asking about the trolley problem and you all went, oh, the trolley problem. People on Turk are doing that even more than you are. They see the trolley problem, they've seen it. I guarantee you, it's very difficult to find a Turker that has not seen the trolley problem. They've seen it. They've seen it 1,000 times. They've seen all the variations. And they complain about it. And they're satiated. And they're sick of it.

They will say things like, if I see one more trolley problem, just kill them all. Is there a way to kill the five and the other person on the other side of the track? I don't care anymore. OK, they're completely satiated. It's kind of like saying, hammer, hammer, hammer, hammer. It loses all meaning at some point. You don't have the gut intuitive response. And even if they're doing their best, even if they're not trying to fool you, even if they're honestly trying to answer, they just can't. They don't have the gut intuitive response anymore for the stuff that you're asking them.

Some ways to get around that is to try even simple changes. Just don't call it the trolley problem anymore. Set up something else, which is not 5 versus 1, which is 10 versus 2 and it involves pineapples. Like something. You know, these small changes can matter a lot. So that's one thing about repeated exposure, that it destroys intuition and related to that, there's something called super Turkers. These are people that, you know, 1% of Turkers is responsible for about 10% to 15% of all studies. So when I say people have seen it a lot, that's part of the reason.

There's a lot of people doing-- the same small group of people is probably responsible for a lot of these studies. The other reason that repeated exposure ruined things for us is because, you know, it just ruins basic correlations. So let me let me give you an example. Who here has heard of the ball and bat question? Who here has not heard of the ball and bat question? OK, let me pose it to you. Those of you who suddenly say, oh yeah, I know this. Sh. I'm interested in people who have not heard this before.

So it's a simple question. A ball and a bat together cost $1.10. The bat costs $1 more than the ball. How much does the ball cost? I'll explain that again. A bat and a ball costs $1.10 together. The bat costs dollar more than the ball. How much does the ball cost? Anyone? Shout it out.

AUDIENCE: $0.05.

TOMER ULLMAN: Yeah, $0.05. Who would have said $0.10, don't be shy. Thank you. Thank you for being brave enough. A lot of people say $0.10. They say immediately. The ball costs $0.10. The bat costs $1. Wait, but it costs $1. No, no, it costs $1 more. So it's-- does everyone get why $0.10 is not the answer, it should be $0.05? OK, good. This was sort of a classic question along with two other questions called the lily pad question and the widget question. The widget question is something like five widget machines make five widgets in five minutes.

How long does it take 100 widget machines to make 100 widgets? Five minutes, right. Not 100. These sort of three questions were found out to correlate a lot better with things like IQ tests or even better, sorry, than IQ on a lot of different things. They've been found, you know, many smart people from MIT and Harvard included have failed this question. And the point was, you know, people have made a big deal about it. Like this is even better than IQ tests on all sorts of measures. And it's so simple. We don't have to run 100 IQ test questions.

Let's just ask the ball and bat question and then see what that correlates with. And it also relates to, you know, Kahneman makes a big deal about it. It's system one versus system two. System one really wants to answer that it costs $0.10. And you could solve it. Of course you could all solve it. It's not hard for you if you wrote down the simple equation, it's trivial. Right, you wrote it down as like x. And x costs this. And you could all solve this. You could all solve this in middle school.

But you don't. Like system to-- some of you do. But usually the first time you hear it, you don't. And you don't-- even if people are warned, like this is a bit of a trick question, you know, think about it, they don't, usually. And people have used this as like ways to test system one-- do people know what I mean when I say system one versus system two? Who doesn't? Raise your hand. So very, very quickly because it's not that relevant to Mechanical Turk, but very, very quickly, you should read *Thinking Fast and Slow* by Kahneman. The point is that the mind can generally be categorized into two systems, even Kahneman doesn't quite believe that, but it's sort of this thing that's easy to talk about.

System one is the fast, heuristic system that gives you the cache response system. Two is the slow, laborious, can't do many things at one time system. That's the sort of thing that you would use to solve algebra problems. That's the reason you need to slow down when you're thinking about something very hard. System one is the sort of thing that gives you biases and

heuristics and things like that. This was given as an example of the bat and ball was like one of the prime examples of system one wants to do this, system two wants to do that. System two is lazy, doesn't actually get engaged unless it really, really has to.

Why am I bring this up? The reason I bring this up is because people thought it was a great idea to put it on Amazon Mechanical Turk and ask people a lot of questions to see what it correlates with. Everybody knows about the ball and bat problem on Amazon Mechanical Turk. Everybody. Don't think that you're being unique. Don't ask them the widget problem. Don't ask them the lily pad problem. They all know about it.

And here it's not a problem for intuitive, you know, being satiated and things like that. Even if they want to, right, I said before, like they want to tell you what the thing is, they just don't have the gut response anymore. Here they just know it. The reason most of you answered, or those of you who knew about it said, haha, $0.05. I know that one. I've solved it before. People on Mechanical Turk know that too. And it's sort of destroyed any sort of measure it had for whatever the heck it was trying to measure.

In general, there's sort of this growing sense that since 2010, people have been trying to make Amazon Mechanical Turk more popular for a few years. It feels almost like an overexploited resource, a little bit. Like you have this tribe which doesn't know about numbers. Let's all go and study them and ask them a billion questions. And the reason they don't know about numbers is because they don't know English. And by the time we are done with them, they will know English. And the one thing they'll know in English is how to count to 10 because we've asked them all these questions.

Amazon Mechanical Turk feels a little bit like this over exploited resource at sometimes. These have been more concerns for you as a requester working on Amazon Mechanical Turk, things to sort of keep in mind and watch out for. Here are some concerns for people on Mechanical Turk that you can try to alleviate. And these sort of things about-- I've already mentioned sort of two of them about low pay and rejecting people for no good reason and things like that. One more thing I want to point out is this thing about no de-briefing. When people coming to the lab, you can tell them why they were in the study. That's a basic part of the protocol of psychology.

When people are on Amazon Mechanical Turk, it's a good thing to put it in the experiment, why were you in this experiment if you have such a thing, and you're running an experiment.

People might drop out in the middle, might decide it's not for me, actually, I'm done. And they never actually figure out what the point of the experiment was. You might say, well, who cares? But you might say, well, who cares to de-briefing people in the lab. If there's a reason for de-briefing people in the lab there's probably a good reason for de-briefing people on Mechanical Turk. If they drop out before their de-brief, that's kind of a problem.

I don't have a good solution for it. It's something to keep in mind. There's also the problem that you should keep in mind and report, probably, if it's a problem, the amount of people who have dropped out in the middle because otherwise it can lead to all sorts of small effects. Like if your task is really hard, and a lot of people drop out, you had like a 90% dropout and then you say, oh, people are brilliant at my task because the 10% who actually stuck through with it are the sort of crazy people who are willing to do it. That's actually really, really skewed. So keep in mind the dropout should be low. It should be like a few percent or something like that.

The flip side, by the way, of the no de-briefing problem is a coercion problem. So when you bring people into the lab and you say, you know, do the study. You can stop at any time. No problem at all let me shut the door and wait over here. There's sort of a slight feeling of coercion, even if the study is not something they really are enjoying doing. They'll still do it because they feel pressured to. That problem doesn't exist, at least on Amazon Mechanical Turk. I mean, they're in it for the pay and there's some cost and all that. But if they really don't want to do it, if it offends them in some way, they'll stop. So that's actually a bonus.

Some general tips. Let's see. And I think I'll have two more slides and then we'll wrap it up. Some general tips. In general, when you're thinking about your task, the lower level it is, the closer level it is to-- think about the Stroop task. OK, let's put the Stroop task in one case and the question I actually asked you in another case. The Stroop task is low level, hard to beat, even if you've seen it a million times, you will still find that effect. If your task is like that, you should expect to replicate it.

If it's much more higher level, the sort of thing that relies on them not having seen that before, it's harder to replicate. You want to make sure that people who see it have not seen it before. If it's high level and relies on some sort of zeitgeist, like what people think about AI right now, in two years you'll find a different result. Don't expect the thing that I put online today to replicate in some sense in two years. Have participants give comments. At the end of your survey, once they're done, before the de-briefing, collect some demographics and leave them something optional to just say, what did you think of this study? Do you have any comments?

Or anything like that.

Most of the time you won't get any comments. After that, the most likely thing is they'll say, fun study. Thanks, if you've done it correctly. Interesting, stuff like that. Some studies we've done have been crazy and they're very, you know, they really like it. And it's nice to get good feedback. Or they'll tell you something like, that was really interesting, could you tell me a bit more. Here's my email address. Or this button that didn't work for me. Or I was actually a bit bothered by the fact that you said that you would kill the robot. Things like that. Or you ask them things like-- give them comments like, why. You don't put that necessarily in your experiment.

But you tell them, like, you know, do this. Do that. Make a decision. Why? Like the trolley problem. Why? OK, it's the sort of thing that's not likely to be published immediately, but it's definitely the sort of thing that will help you think of where to take your experiment next. It's very, very important to communicate. And what I mean by that is give them an email at the beginning to reach you in some way, say like, you know, in the consent statement. You're doing this experiment. You're doing it for x. Here's a way to reach x if you want to.

AUDIENCE: Do you give them your actual email?

TOMER ULLMAN: You can set up a bogus email and the sense of, like, tomer@mechanicalturk. I personally give them my email at MIT. I do. And make sure that you respond to their concerns. And they will write to you, especially if something goes technically wrong. They'll say, like, you know, the screen didn't load for me. I forgot to paste in the code that you wanted, things like that. Or this didn't work quite well. Write back to them. Explain what happened. If they want to know what the study is about, you should explain to them what this study is about.

You should do that for two reasons. It feels silly to mention this. I'm sure you're all, you know, you've figured it out by yourselves. But I'll mention it anyway, just on the chance that there's one person that says, oh yeah, that's a good reason. For two reasons, one is that they'll like you a lot more. OK, these people, they go to their own forums. There's a lot of Mechanical Turk forums. There's hundreds of them. And they tell each other what things they should do.

They do a good job of policing each other. They're like, they try never to post answers to things, or like, oh you can do this by answering this question. No. It's like, they don't tolerate that because they know that we don't tolerate that. You want them to like you in that sense. You want to get good reviews on these things. And one way to garner good favor is to

communicate. The other reason is because it's just a good idea. These are people in the public. You wouldn't think about not answering a question of someone who came into the lab and asked you something.

Keep in mind, there's a real people behind the screen. Make sure that you treat them as real people. I don't mean-- I sound like I'm berating you or something, like that you guys have not been communicating and it's awful. No. That's not the point. I'm sure you all mean to do that I'm just trying to emphasize it. Like I said before, don't reject unless it's an extreme situation. Also, decide ahead of time how you're going to reject. Decide ahead of time on a catch question, something I'll get to in a second. And say, I'm going to reject people if they do this and stick to it. Because otherwise, if you don't do that, then when it comes time to actually try to write a paper you'll say, well, I think I'll try throwing out all the people that did it in under 20 seconds because I don't think they were paying attention that much.

Maybe 30 seconds. Yeah, this test should really take 40 seconds. You got the point. Decide ahead of time on the rejection criteria. Have good catch questions. This is good both for, you know, knowing who to reject and making sure that they're paying attention. Catch questions are the sort of thing that you would put in the middle of your survey or at the end of survey, or at the start of the survey, just to make sure that they are paying attention. Ideally it would be also that they have actually read the instructions and know what they're supposed to be doing.

So sometimes even if they're paying attention, they didn't get the instructions or something like that. There's a bunch of different ways of doing this. One way of doing it, you know, I'm sure you guys can come up with your own ways I'm just giving you some examples of the stuff that people I know have done. Toby, for example, that you've seen doing some counterfactual stuff maybe, he just gives people a screen with the instructions and asks them some questions. And until they get the question right, they don't move onto the next screen.

So he doesn't reject them later. He just says, in order to pass to the next screen, you have to answer this question correctly. And he has some way of checking that. That's been really good. Like once he implemented that, the data is much, much better and cleaner. Here's the sort of catch questions you don't necessarily want to do. They're very popular. You don't necessarily want to do them. They're things like, have you ever had a fatal heart attack. And the answer is, of course, no.

Have you ever eaten a sandwich on Mars. It's the sort of thing that like you're trying to catch

people that are going through it very quickly and are just marking things randomly. One of the reasons you don't want to do that is because even if they're answering randomly yes or no, you'll still miss 50% who just got it right by error. The other reason is the standard stuff. I mean, I'm sure you guys could come up with something. But there's a lot of examples out there. The two examples I just gave you, the Martian, the fatal heart attack, this is stuff that gets used over and over and over again. And they sort of just know it.

One of them said like, any time I see the person I told you before was like juggling kids and trying to answer at the same time. He says, oh yeah, whenever I see the word vacuum, I know it's time for an attention check because it's going to be like, have you ever eaten a sandwich in a vacuum or like something like that. But whenever I see vacuum, it's obviously an attention check. You don't want to do that. Ideally, you want to have something that relates to the task.

So in one of our examples, we were doing some sort of Turing task. And we just wanted to say, like, here, complete the following sentence. You were playing the game against a-- and then it's an open text box. OK, some of these people have like automatic robots that fill it in. So they'll do something like, thanks. Or yes. Or something like that. Then they just hope that yes will match. But here the correct answer was robot. You were playing the game against a robot, or against a human, or something like that.

Did people get that example? OK. So ideally, the good catch question is an open field, something that's not just you can click and get right by mistake and relates something to the instructions that you were giving. This is not a tip. This is something you should do. Again, it's trivial. You'll have to do it if you're thinking of running it in your own university and your university has never done Mechanical Turk before, get IRB approval specifically for Mechanical Turk. So just make sure you get IRB approval.

Make sure you get informed consent at the beginning of your study, say like, we're going to be doing this. If it's OK click on this button that says, I agree. Usually the IRB will force you to do that anyway. And as I said, ethical pay, I just keep going back to that. OK, there's various helpful tools for running experiments. If any of you are interested in this, reading more about it, much more in depth how to actually run an experiment and things like that, come talk to me afterwards. Or look at Todd Gureckis's website. Other websites that you should check out.

These are the forums you should probably know about, TurkerNation, mTurkGrind, and TurkOpticon. These are useful both to get a sense of how your task is doing, are people sort

of responding to it saying something about it. It's also a good place to publicize your study. If I need 300 participants within two hours, I can put it on Turk and hope the pay is enough. Or I can put it on mTurkGrind. People there have liked our tasks before and they'll give you a thumbs up and you'll get 200 people within an hour just because they know about you and they know that you're an OK person, and you communicate.

So it's a good practice to get a user name on one of these things, or both of them, when you run an actual experiment. Explain what you're doing. Put it on there. Be willing to answer questions. TurkOpticon is the usual thing that will then be used-- this is what one of these forums looks like, by the way. They're like, you know, oh it's terrible. It's Tuesday. What should we do today? Here's how much I've done today. Somebody says, I didn't see this posted anywhere. Very interesting to those of you who want to do 3D printing. Then there's like the title of the experiment and a description of it. And people can just click directly from there to the experiment.

Usually the experiments that they post on the forums looks something like this. They're getting these numbers off of that last website that I just mentioned, TurkOpticon. So usually when people want to rate you on Mechanical Turk, they're not going to complain to Amazon because Amazon's not going to do anything about it. Like I said before, requesters have a lot of power and Amazon doesn't bother to arbitrate, usually. So one of the ways that they do have of either rewarding or punishing you is to go to TurkOpticon and give you a rating for communication, generosity, fairness, and promptness.

And those numbers will then be used on most of the forums when you publish your task. People can see, like, these guys have actually cheated people before, or something like that. So you can go, once you've registered, go to TurkOpticon and you can check out your own ratings and things like that. So yeah, in summary, Mechanical Turk is a wonderful resource. I hope I haven't scared you away from it, those of you who are thinking about doing it. It amplifies everything. You can do a lot more a lot faster. But it doesn't get rid of concerns, it amplifies those concerns.

So things like ethical concerns and payment concerns, and things like that, the sort of same sort of concerns that you would have in the lab, those are included there, and other magnified 100-fold because you're recruiting a lot more people.