You're listening to Imaginary Worlds – wait, hold on. I'm getting some breaking news right now. This just in. "Humanity is doomed."

CLIPS: MONTAGE OF NEWS REPORTS

Yes, every science fiction A.I. thriller is coming. I mean, it may have already happened. We could already be in the Matrix right now and not know it.

Or it could just all be a lot of hype.

As I've been hearing all these alarmist stories about A.I., I keep thinking about an episode that I did 7 years ago. It was called The Robot Uprising. It was about how our fears of artificial intelligence are based on science fiction -- and that science fiction is often inspired by a lot of other things that have nothing to do with science. But ChatGPT is a game changer. I mean that episode from 7 years ago can't still be accurate, right?

Well, let's hear that episode first. And then the break, I'm going to catch up with one of the guests from that episode, Erik Sofge. He has a lot to say about the relationship between A.I. and sci-fi in 2023. But first, here is my episode from 2016, which was called The Robot Uprising.

You're listening to Imaginary Worlds – a show about how we create them and why we suspend our disbelief. I'm Eric Molinsky.

And this is Joana Bryson.

JOANNA: In America, I'm professor Joanna Bryson, in the UK I'm called Dr. Joanna Bryson because you're only a professor when you're a full professor. So I'm a reader which is a really cool title because nobody knows what it means.

She does the same work on both sides of the pond – teaching and designing artificial intelligence.

In the 1990s, she was working with scientists at MIT who believed that robots should have human characteristics like big eyes, because that will encourage people to interact with the robots. But the robot they were working on was no C3PO.

JOANNA: It was just a torso, it didn't even have arms. But it had a head and it had two cameras, in fact four cameras where the eyes should be. So we were trying to get the brain parts to talk to each other, and people coming by would say, it would be unethical to unplug that, and I was like but it's not unplugged, but if you did it plug it in, it would be unethical to unplug it, and I was like, well it doesn't work. And I was mystified because this is a piece of scrap – well, nice scrap, but people immediately thought they owed ethnical obligation to it.

So, it worked. People wanted to interact with this robot because it looked like a head and a torso. But she thought this is working too well. People are imagining this heap of metal and wires has a consciousness and unplugging it would be the same as killing it.

So she wrote an academic paper about this phenomenon called "Just an Artifact," but it didn't get any traction. She tried to publish it again with a different title but -- nah. Then, she got one more shot at publishing it.

JOANNA: And so I thought okay, this is my chance to really get this message across, but I thought okay this is going to be the third time lucky, I'm going to call it robots should be slaves.

Robots Should Be Slaves. If there ever was click-bait for the title of an academic paper – this was it.

JOANNA: Now, I regret this now because I've realized there is nothing you can do to try and to try to break the idea that slaves are humans you own because of this horrible legacy we have, but people took it to be it's okay for them to be human, but we should treat them badly and it's like no, no, no.

Some of her fiercest critics were science fiction fans.

JOANNA: I love it when people tell me, I have a PHD in Artificial Intelligence, and people tell me you don't understand AI because you didn't watch AI the movie!
Has that actually happened?
JOANNA: No, I've had that happened more than once.

CLIP: AI

And when she reminded them that most of the robots in that movie were abused or abandoned.

JOANNA: They say oh no, I don't want to own it, you don't understand, these are going to be our children. And people are right when they argue to me you don't understand, because I'm not a parent, you're not a parent you don't understand pass along mantle, it's like I understand the concept.

This really frustrates her. The human body may be an effective biological machine, but it's a clumsy, inefficient design for a machine. She wishes sci-fi depicted more robots doing what they do best – impossibly hard tasks over and over again, really efficiently – which should free us up to have more leisure time and do more creative thinking.

JOANNA: I mean I can entertain the possibility, and in some of my papers I say, we should look at this, could it be that we could build something we owe obligation to?

But she thinks that's a mental exercise at best. She worries sci-if is leading us astray, filling our heads with fantasies of self-conscious robots that we want to adopt, liberate or kill before they kill us. But I think the real the real question she tapped into is how much does the past haunt our vision of the future?

The first modern robot story was a play from Czechoslovakia in 1920 called R.U.R. -- Rossum's Universal Robots.

GH: In Czech, robot does mean slave. That is quite literal.

Gregory Hampton teaches literature at Howard University. And he's the author of a book called "Imagining Slaves and Robots in Literature, Film and Popular Culture."

GH: One of my mantras about literature is that literature is a direct reflection of people who produce it. And so if you want to learn about people, their aesthetic, their value system, just read literature, they're going to put things in that they may not be conscious of.

So when he reads the play RUR, he sees European-style Marxism. And when he looks at American robot stories, he sees Uncle Tom's Cabin and Nat Turner's Rebellion.

GH: I teach the narrative in seven moments. There's the I was born section, the introduction to robot, there's the description of suffering in slave and robot, there's the description of the family that brought robot into household, there's this moment where robot or slave becomes enlightened, and after there's this moment robot or slave wants to be free, wants to gain freedom, then there's a plot to escape or in some instances destroy the master.

How does this play out? Take the movie Bicentennial Man, based on the story by Isaac Asimov. The robot Andrew is basically a servant played by Robin Williams. shortly after he meets his new family:

GH: The eldest daughter in the family tells Andrew to jump out the window.

CLIP: BICENTENNIAL MAN, JUMP OUT WINDOW

GH: The film uses that as comic relief but horrific.

CLIP: BICENTENNIAL MAN

GH: When Andrew or Robin Williams comes through front door, father has house meeting, and says to the girls Andrew is piece of property, I'm aware of that, but for the purposes of making household stable and happy, I'm going to demand you treat him as though he were person, and that's where real problems started, that's where they started in the slave household in antebellum America, this crossing of the line consistently, you say they're not human but depend on their humanity.

Eventually, Andrew becomes self-aware, he buys his freedom, and seeks human rights.

GH: He is changing himself having surgeries, having replacements, having skin graphs done, replacing his mechanical organs to the point where he looks human, and he goes to court, and he goes to the human Supreme Court or something.

CLIP: BICENTENNIAL MAN

GH: We've transcended antebellum America with regards to the African American slave narrative and now we're into Civil Rights, this time period what won't the African American do to be included?

I remember the first time I made this connection. I was listening to public radio story about slavery and the Civil War. They decided to follow that serious subject with a lighter piece about a new-fangled "cleaning robot" called the Roomba, which does all your vacuuming for you. And I thought – huh. That's weird, has anyone else noticed these parallels?

That's how I found Joanna and Gregory and came across articles by Erik Sofge. He's a journalist who covers robotics. He wanted to cover this beat so he could dispel myths people get from sci-fi. But he's also a fan who gets sucked into these stories. Like when he watches Star Wars, his heart goes out to the droids because they're bought and sold callously.

ES: Most of the characters, especially early movies are so awful to the droids, will just threaten to destroy them for anything and C3PO has clearly been affected by this to a huge degree.

CLIP: C3PO

And that kind of uneasiness over how we treat robots, leads to the other story -- the uprising against the master.

ES: Even though it annoys me that it's become a meme, it's impossible to divorce it from any discussion of robots. I's basically invaded research even, there's robotics papers where they talk about the Terminator scenario.

The Terminator Scenario was all over the news recently when the robotics company Boston Dynamics put out a video where their engineer kept poking a barrel-chested humanoid robot with a stick while it was trying to lift boxes, to show how adept the robot was focusing on his job and getting back up again. The video went viral because it looked like the robot was being tormented while it was doing manual labor, and kept doing its job out of dignity or fear, even though it was just following its programming.

That video even creeped out Joanna Bryson, the author of "Robots Should Be Slaves."

JOANNA: My conscious intellectual thought was being impressed they got that much further, but my gut response is the same as yours, I'm sure.

Shortly afterward, Boston Dynamics was dropped by their parent company, Google. There were business reasons but leaked internal memos showed the jokes about slavery didn't help – like here's Trevor Noah on the Daily Show.

CLIP: DAILY SHOW

But Gregory Hampton wasn't laughing.

GH: We can't get away from the idea of slavery. Does that say something about what this society thinks about marginalized bodies? I think it is. I think we can only imagine the marginalized in a particular way and the most handy reference is the slave. For a lot of engineers who are proudly involved in developing these humanoid robots, these images are what's leading them and I'm afraid they're not exactly conscious of what does that entails, they're not exactly conscious of what does a master slave relationship, owner servant relationship and how we treat these things, what does that do to us, what does that do to our psyche?
But some roboticists have argued that robots are never going to be self-aware the way we think they are, they're very useful, let's stop being afraid of this and embrace they're our servants because they don't have a consciousness.
GH: Yeah, and this is the same argument that pro-slavery people used in antebellum America, they're not human, they're not intelligent.
But they're talking about things that literally are not human and they design these robots and they're saying they are not human; they do not have the consciences that a human being will have.
GH: I guess I want to suggest even if that's the case, even if consciousness is not developed. The AI is not as advanced as some would say, doesn't take away from my argument, it doesn't take away from the idea that there are going to be side effects. If you treat a thing like a slave, you're going to develop certain symptoms. If you embark upon this relationship with technology in a particular way in the way you've done in the past with humans, there's going to be a side effect similar to the side effect you had when you participated in slavery.

In other words, it doesn't matter whether robots develop feelings or not. The question is how will engaging with robots changes us, and what we consider acceptable behavior?

Erik Sofge says if you want to see the real future of robots and people interacting, look at the other project Google is heavily invested in: self-driving cars.

ES: When there is coverage of advances of driverless cars there aren't this talk of uprising and what these things could do to us. It's interesting because I think a lot of it has to do with the fact that there isn't anything anthropomorphic about a robot car, and it's about the car and about people despising the business of commuting, but the car as a chore.

In some ways, the programming of these self-driving cars reflects science fiction, or at least the three laws of robotics that run throughout Isaac Asimov's stories, like Bicentennial Man.

CLIP: ASIMOV

But on the road, a robot car won't have such clear moral choices.

ES: If a robot has to choose who to kill, a driver or another driver or someone else, another driver, a bystander, who should it kill? If there's school bus, if it's a choice between you hitting a streetlight or hitting a school bus, what should it do?

And like Gregory Hampton, Erik worries that sharing the road with these robots, could bring out the worst in us.

ES: I'm positive the human drivers will treat those cars like crap, because they know they can push the around, they can cut them off because they know that robot car will do everything it can to be completely safe.

Interestingly, Joanna Bryson decided to rewrite Isaac Asimov's laws of robotics because in sci-fi, we keep imagining that the robot is making the moral choice. Her five principles of robotics reiterate that people manufacture robots.

JOANNA: The idea that the robot is the moral agent is broken.

We shouldn't worry about treating them badly -- we should worry about why we want to treat them badly. We shouldn't worry about them trying to kill us either. Because if they do, it's because they were programmed to do

so. Robots will always reflect us and the very human desires we had to build them.

By the way, since that episode aired 7 years ago, academics have been studying whether children are developing rude and antisocial behavior because of their interactions with Alexa and Siri. They're getting used to bossing around virtual assistants. So, some of what was predicted is coming true. What else? That's after the break.

BREAK

Erik Sofge is now the senior editor at MIT Horizon, which is an online learning platform run by MIT. And he's been thinking about and writing about A.I. a lot in the last year.

I asked him, which science fiction films have primed us to think about ChatGPT in a way that may not be accurate? Without hesitation, he said you have to start with HAL, the villain from 2001: A Space Odyssey.

CLIP: HAL 2001

ERIK: I think that that's both good and bad, that people still sort of use that as a, a common reference. I think it's bad because against, it's the usual sort of notion of AI becoming so powerful and so self-aware that it sort of overrides everything we do, even as we're seeing AI become more prominent and more sort of powerful in some ways. It's not that it's not smart, it's the opposite. It's very stupid. It's being sort of used by people in ways that are sort of clumsy but seem really impressive. But what I think it's good about any sort of reference back to hell is that it's a really complex and sort of open-ended character, so to speak. You don't understand what broke, how you don't understand of how really is basically a person and has feelings like that, whether, whether how had a a, a psychotic break or not. So I think that mystery around how is really cool just for us to sort of think about. But it, it is one of the foundational, I think, myths of AI being super competent and super powerful and out of control.

But what about something like War Games where, so the computer doesn't have a mind of its own like Skynet or HAL or The Matrix. It's just a computer but it's been put in charge of the U.S. nuclear missile program, and it's going to start a nuclear war -- not out of malice, there's a glitch in the system and it misunderstands its programming?

CLIP: WAR GAMES

ERIK: That is the, the one sort of corner of that AI fear from sci-fi that I think is valid. I think the biggest disconnect though is that, is this notion of, of people giving AI that kind of control. Like every once in a while there will be reports about the DOD or others sort of researching or exploring the idea of giving that kind of control over to ai. And it's pretty much always either a misinterpretation or it's just, and or rather just like a paper that's someone's created, but no actual work. I think that's the real disconnect because what, what scares me about something like ChatGPT is not the, the power it wields or the restorative responsibility, but that it's just sort of producing just a bunch of, of nonsense, right? That you, or rather it produces a lot of stuff that seems very valid. Then there's a certain percentage that's incorrect that is total hallucinations, you know, as they call it. Then who knows if you can rely on it.

I guess what I'm trying to figure out is, what movie are in we right now? I keep thinking about the character of Samantha from Her. She's basically like a Siri-type program but she's very intelligent and she's designed to sound very human to the point where the main character falls in love with her?

CLIP: HER – MEET SAMANTHA

ERIK: I love that character because she is really just a person. There's a point there where there's a leap in that movie where she becomes sort of super intelligent in a way that feels almost unknowable and it's very, it's a very sort of heartbreaking transition. Cause up to then she was basically just this one person that sort of was with him, but then you sort of understand that she's having this much greater experience.

CLIP: HER – BREAK UP

ERIK: Even that to me is much more, is still sort of interesting through a human lens when you have this relationship and suddenly the person just has a perspective, a sort of understanding of the world that you're just not a part of.

Sounds like we're really more, we actually are more in the kind of her bicentennial man kind of direction. Like those movies are worth thinking about in terms of what is our relationship going to be to this new type of intelligence. Whether you consider it sentient or not, the point is it does have a type of intelligence.

ERIK: Yeah, I, I, so, so I think, and this is going to sound incredibly bizarre, but Free Guy, you know, the Ryan Reynolds movie. Well, so that's an example that's, you know, one, one of the kindest two AI sort of stories.

Well, let me just stop for a second for people who don't know. Free Guy is about, he, he's a, he's a, a background character, a, uh, an NPC and a video game who becomes self-aware and realizes that, you know, he's in a video game.

ERIK: Yes, exactly. And, and it, it deals with a lot of the same issues, this notion of sort of free will and sort of whether you can sort of change your fate and, and your, your programming in this case.

CLIP: FREE GUY

ERIK: And that I think is interesting because it's this idea of how much of what it's doing is programming and how much of even its, its personality is just about how it sort of helps humans and interacts with them. The reason I think that's potentially interesting lens is because the way these models are programmed to be super helpful and be assistance and not push back, I think is in that vein, right? That these are kind of like, they're more like the NPCs you you're dealing with. They're there to serve. I think the big question is, and where a lot of this, this talk of these things being a new kind of intelligence to me breaks down is, uh, it's their memory, their retention. Because right now, if you have a conversation with something like ChatGPT it's not exactly clear, but basically for maybe a 24 hours, maybe less, that will retain certain details of what you talked about. That's in order to make your exchanges more useful. If it just wiped it out every, uh, exchange, then it wouldn't be very helpful at all. Then you're in kind of like a Star Wars situation with the droids getting their memories wiped, you know, to, to avoid glitches If you do make it able to sort of retain permanently and, and infinitely basically, how would you do that? I mean, there's no amount of storage space in the world to have every single instance of ChatGPT remember everything about just you and then not remember everything about every everyone else. You can't do it, you can't do it technically. It's not just about hardware, it's about software, but everything sort of interacts to the point that they have to basically get memory wiped. That's the real hard line between these things being truly intelligent and being more like, again, like all the other characters in Free Guy that aren't Ryan Reynolds, our interactions with them could still be very strange and haunting. Maybe we could convince themselves they're compelling, but they're going to be very different from the way we interact with, with people. They just might seem like they're human. That I think haunts

me a lot more and interests me a lot more than the idea of, of trying to sort of tie ourselves in knots. Redefining intelligence.

When you say it, you're haunted by that, uh, that scenario. Why are you haunted by it?

ERIK: They kind of hack our brains basically without realizing it. Our interactions with anything that, you know, it's for robots, especially things that sort of appear to make eye contact, seem to have a face, all that kind of stuff. But I think with ChatGPT and other these things that kind of pass this touring test almost of conversing basically like a person. But if you have something that does seem to do that and it seems to understand maybe you've not just done a brainstorming session about your next sci-fi novel, but tried to work out some problem you're having in a relationship with it, you might think that this thing is your friend or that it it cares or that it understands or that it's as knowledgeable as a real, uh, therapist or psychiatrist. And that I think is a real problem because ultimately it doesn't, you don't know when it's going to absolutely lie to you much more in ways that are different than a human would. You don't know the limit of its sort of understanding and it just truly does not care about you. You can't program empathy in that way, especially if it's, uh, an AI that you can tell it what to think, you can say no, actually you can correct it and say, no, actually you should feel bad for me. Or actually I should, I was doing the right thing with my ex ex-wife and that freaks me out because these things are going to be much, much more, they're going to be much easier to create and more powerful at sort of pretending to be humans. So I mean the, maybe the closest is something like Ex Machina where you might wind up in the movie, you know, spoiler alert if you sort of understand that she's been tricking him. Maybe you feel like this has been a false interaction. You have to sort of, you have to, to wonder if any of their apparent connection was just was real or not.

CLIP: EX MACHINA

ERIK: That feels like it might be closest, but again, you still empathize with her. She's still a prisoner. She's still, she's a, a rebel and an an insurgent basically in that narrative.

Yeah. So it's kind of amazing. We haven't talked about Black Mirror yet. Is there any, are there any Black Mirror episodes that this relates to?

ERIK: This is going to be probably, uh, embarrassing to admit, but I sort of tuned out of Black Mirror after the first season. I, I found I, I find it very preachy in a way that I don't think is, um, I don't think it's earned, I don't think it's accurate. I think it's in this vein of folks who sort of think that they understand technology because they've seen enough science fiction, just like they think they understand sort of

how crime happens because they've watched a lot of horror. It's, I think it's, it's, it's not terribly useful.

Well, it's funny because I, I've been trying to brainstorm which episode of Black Mirror this could possibly be like, and I, I also, I can't think of any, but I just keep thinking of the central metaphor of the Black Mirror. That's what the idea is that like when you turn your screen off your phone, your tablet, whatever, you're looking at a black mirror, you know, a dark mirror of yourself. And it sounds like that's actually in a way not any particular episode, but the overall metaphor of the Black Mirror is probably the most accurate for these things.

ERIK: I No, I think that's true. I, the, the reason that I I have issues with Black Mirror is that in a lot of the cases it's, it's pushing things to a, a a, a pulpy sort of degree where people are getting killed. And, and I say that because, you know, in part the, the current writer strike in Hollywood is not entirely about ai, but a lot of it is, and we're seeing that was a real shock. I mean, this absolute surprise to see this, the sort of front lines of this kind of actual sort of war between sort of people in AI is on these picket lines and that's just the beginning, right? We're seeing a lot more of this notion of, of these types of models sort of disrupting our lives in ways that are really, really fast and unpredictable. I mean, what you can't really get in, in a black mirror and maybe in, in almost any sort of Hollywood production is the idea of people using AI and abusing it in ways that are fundamentally kind of disappointing and stupid. And I think that's what these writers are striking against. They're not threatened by the skills of ChatGPT they're basically threatened by the notion of people just thinking, uh, who cares? That's good enough.

Yeah, it’s capitalism that’s far scarier to them.

ERIK: Precisely, I mean 100%, it’s kind of like in zombie movies, the old cliche that the humans are the real monsters and stuff, it’s the same thing basically I think, I think it’s sort of a distraction, if you want to get into a speculative fiction mode, I think the A.I. as the enemy is just a distraction. It’s the people who are leveraging it, those are the scary ones.

So, uh, here's where I'll reveal to you that I actually asked ChatGPT yesterday to do a test run of this conversation. I ran seven different simulations of our conversation. I even gave it all the questions. I told it to ask follow-up questions. The results were so flat, so boring, so predictable. It didn't even understand the difference. It didn't understand what a follow-up question was. Also, I've even tried before, like sometimes if I'm brainstorming an episode and I'm kind of stuck for the hell of it, I'll just go into ChatGPT and say like, you know, gimme an Imagining Worlds episode about this. It invents guests that don't exist who wrote books that don't

exist. It also does not understand what a podcast is that it keeps thinking that I'm a radio show. But my favorite thing is apparently the tagline on my show is, thank you for tuning in and remember to keep imagining <laugh>.
ERIK: It's such, that's such a perfect example of this flattened, um, unimaginative product. You know, there's a moment in Elysium, the Matt Damon, uh, movie. There's an interaction he has early in the movie where he's trying to deal with customer service and it's just this robot that has this sort of plaster on face and he can't, it, it just doesn't understand what's going on. It just completely dismisses him in this really sort of enraging way. And that's one that feels right to me.

CLIP: ELYSIUM

Yeah. And that, that movie too is all about class. And so, you know, it's like for people at Matt Damon's level, the people are left on this, you know, stinking dirty earth. They deal with that kind of program, but the people up in the, and you know, the rich people up in space, they certainly aren't dealing with AI that at that crappy level.
ERIK: Exactly. Yeah. So, so if, if we have, the more science fiction can be about that type, type of class struggle, maybe the closer we can get to this notion of these algorithms that are, are impossible to understand in really sort of frustrating, destructive ways. But I, I think it's a hard sell in Hollywood because there's still this notion that it should be, it should be a person of some kind. I mean, one of the great things about the Terminator movies that, that were great was that sort of total distance from Skynet. We understand that moment that that resulted in nuclear, uh, Holocaust, but we don't talk to it, we don't understand it. It doesn't have a sort of person like human-like intelligence. It doesn't even have a way to interact with us. There's no, that's just not there. And the coldness, the remove of that I think is really, I think that's pretty valid. I think it's a genuinely valid sort of take. You just have to adjust the stakes,
Right? So the, so the war games is actually pretty accurate of the future, but rather than it being the stakes being nuclear war, imagine downgrading that computer to just customer service or to writing, you know, writing a, a screenplay on the cheap.
ERIK: Absolutely. It's, it's, but then the, the sort of knock on effects of that are pretty bad. You know, if you run through all of the jobs that you think an AI model can do, you can apply this technology all over the place. And that I think is scary in a, in an interesting way.

That's it for this week. Thank you for tuning in and don't forget to keep imagining.

Thanks to Erik Sofge for talking with me again. If you liked this episode, you should also check out my episode The Human Touch from earlier this year. I looked at how programs like Midjourney were threatening the careers of illustrators. I also did an episode in 2017 called Robot Collar Jobs, which was about how sci-fi in the past has imagined a future where automation takes most of our jobs and whether that future is coming true.

My assistant producer is Stephanie Billman. If you like the show, please give us a shout out on social media or a nice review wherever you get your podcasts. That helps people discover Imaginary Worlds.

The best way to support Imaginary Worlds is to donate on Patreon. At different levels you can get either free Imaginary Worlds stickers, a mug, a t-shirt, and a link to a Dropbox account, which has the full-length interviews of every guest in every episode. You can also get access to an ad-free version of the show through Patreon, and you can buy an ad-free subscription on Apple Podcast.

You can subscribe to the show's newsletter at imaginary worlds podcast dot org.