Legal Status Criteria for AI: Let's not give rights to malware

Miranda Mowbray
University of Bristol
miranda.mowbray@bristol.ac.uk

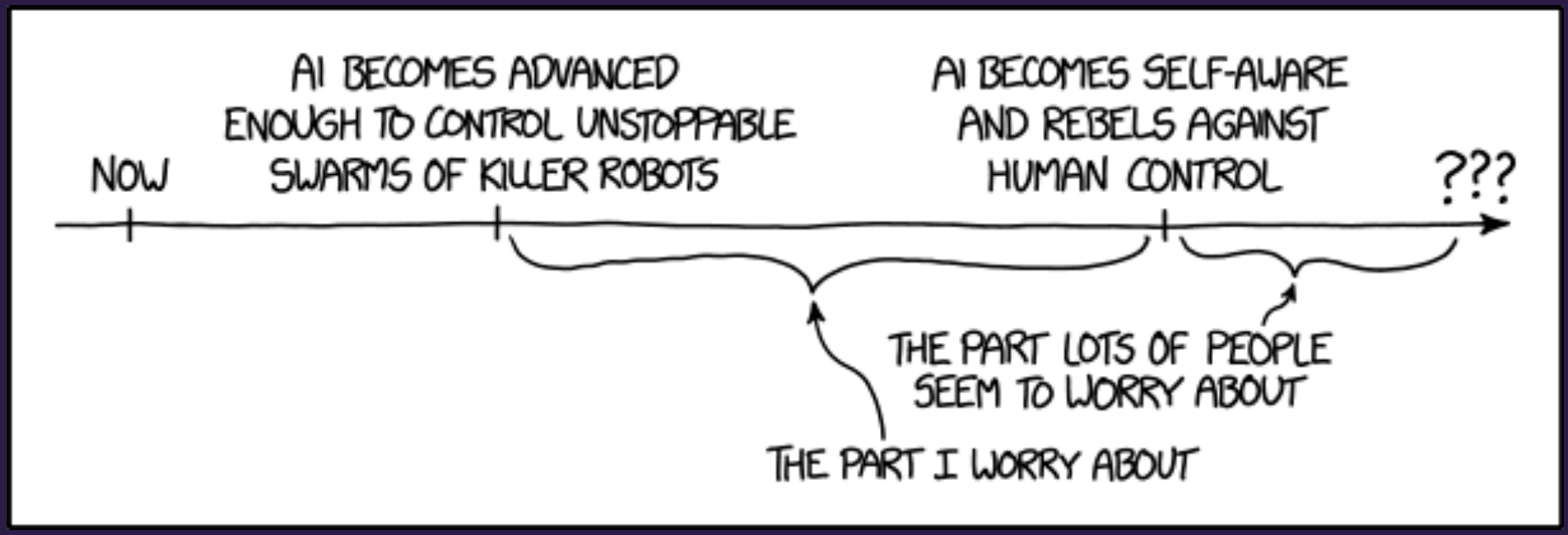
Context

AI / software only
I Am Not A Philosopher
Intended as a provocation

Main argument:
basing rights for AI on some consciousness criteria may $\Rightarrow$ giving rights to malware

I mean malware used as a tool by a **human** owner

xkcd: robot future



https://xkcd.com/1968/

Rights for Malware!

Software has directed evolution
Software is (relatively) cheap

=> If there is a way to make money by exploiting a legal loophole in the definition of AI consciousness,
It **will** be exploited

Malware with rights (to life? reproduction? freedom?)
Malware owners without legal / fiscal responsibility

So what?

Human criminals have human rights
But, malware might become the ***most common*** form of legally-conscious AI

So what?

Malware might become the ***most common*** form of legally-conscious AI

Previously:

- spam (85% of email, March'18 - Cisco)
- ticket bots (Wiseguy, 1.5m tickets 2001-2010 - Motherboard)
- patent trolls ("Most patent cases are now by trolls" Harvard Biz Review 2014)

https://www.talosintelligence.com/reputation_center/email_rep
https://motherboard.vice.com/en_us/article/mgxqb8/the-man-who-broke-ticketmaster
https://hbr.org/2014/07/the-evidence-is-in-patent-trolls-do-hurt-innovation

Spectrum

Consciousness is a spectrum in humans
- Sense & react to environment **pre-birth**
- Mirror test **18 months**
- Sally-Anne test **4 years / never**
- Turing Test **16 hours per 24 hours**

How is AI doing?

Turing Test

Human: my name is mario
Mitsuku: Your name is Mario.
Human: name a country that starts with the 3rd letter of my name
Mitsuku: Russia starts with a letter R.
Human: no bot can do that, If your not human, I'll eat my shirt.

Human: How do you feel?
Mitsuku: I feel with my hands! I feel fine thanks for asking.
Human: You have hands?
Mitsuku: Well unless you think I am hitting the keyboard with my nose…

Mitsuku by Steve Worswick, pandorabots.com
https://twitter.com/MitsukuChatbot/status/943957821774815232
https://twitter.com/MitsukuChatbot/status/955928580034310144

Mirror Test

QBO robot in front of a mirror

Human: Who is this?
QBO robot: Let me see. Interesting. It looks like a QBO. Let me check who it is …[flashes nose light]… Oh. It is me. I must have a mirror in front of me.

QBO robot by thecorpora.com,
https://www.youtube.com/watch?v=EIxoiLmy5mM

3-robot mirror test: Zeng, Zhao & Bai,
https://link.springer.com/chapter/10.1007/978-3-319-49685-6_2
https://www.youtube.com/watch?v=7W5pvbMlOfk

Consciousness: some criteria

Subjective experience

Sensing & reacting to environment - includes online environment

Reasoning & planning - executes logical deductions; puts resources in place for later use; takes different options to achieve same task in different environmental conditions

Internal self-representation - has access to info about its own current state (or state history) as data

Complexity – network size/topology

Unpredictability – does things its programmer can't predict

Turing Test – mistaken for human in online conversation

Autonomy – moves from site to site without direct human assistance; complex automated behaviour over length of time in unpredictable environment

Subjective experience

What is it like to be a bot?
(Apologies to Thomas Nagel)

Malware properties

Sensing & reacting to environment √

Gains info on system, adjusts attack
- Inactive in sandbox, or machine with Russian keyboard
- Exploration
- Ransomware pricing (deadline, geography, file type)

Sandbox evasion e.g. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.363.6295&rep=rep1&type=pdf
Keyboard e.g. https://www.f-secure.com/v-descs/migmaf.shtml
Variable Pricing https://blog.barkly.com/spora-ransomware-variable-pricing-payment-portal

Malware properties

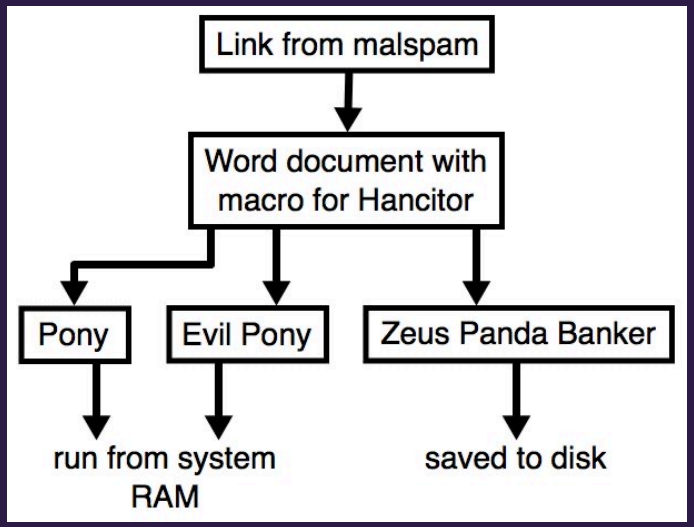
Sensing & reacting to environment √
Reasoning & planning √

Code => execution of logical deductions

Malware properties

Sensing & reacting to environment √
Reasoning & planning √

Code flow => planning



Brad Duncan, malware-traffic-analysis.net,
https://twitter.com/malware_traffic/status/973614618525433856

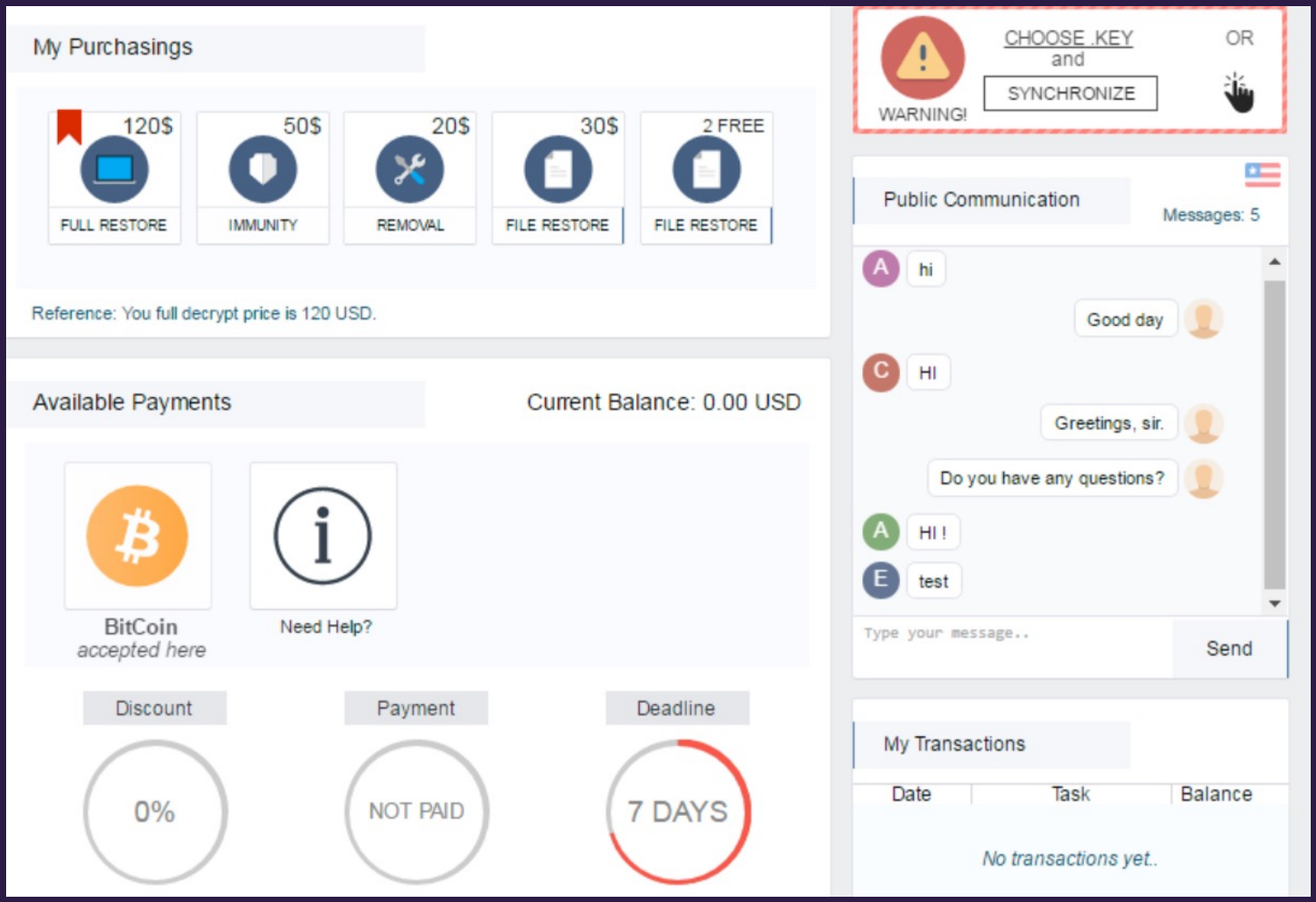
Malware properties

Sensing & reacting to environment √
Reasoning & planning √
Internal self-representation √

```
print(`[+] Sending %d forged IP packets to: %s’
% (power, result[`ip_str’])))
…
print('[•] Task complete! Exiting Platform. Have a wonderful day.’)
```

Memcrashed code, posted by @037 … somewhere

Spora: victim's dashboard



Via Brad Duncan / @malware_traffic / malware-traffic-analysis.net, http://malware-traffic-analysis.net/2017/01/17/index2.html

Malware properties

Sensing & reacting to environment √
Reasoning & planning √
Internal self-representation √
Complexity √

Necurs botnet: Aug-Nov ’17, 1.2m IP addresses in over 200 countries/territories

Jaeson Schultz, https://blogs.cisco.com/security/talos/the-many-tentacles-of-the-necurs-botnet

Malware properties

Sensing & reacting to environment √
Reasoning & planning √
Internal self-representation √
Complexity √
Unpredictability √√

Polymorphic malware
e.g. Sinowal: random number seeder based on twitter trend data

SophosLabs, https://nakedsecurity.sophos.com/2009/07/12/surge-sinowal-distribution/

Malware properties

Sensing & reacting to environment √
Reasoning & planning √
Internal self-representation √
Complexity √
Unpredictability √ √
Turing Test √ √

Turing test

https://twitter.com/brytonbenson/status/970156217283657728

Malware properties

Sensing & reacting to environment √
Reasoning & planning √
Internal self-representation √
Complexity √
Unpredictability √√
Turing Test √ √
Autonomy √ √

Spreads autonomously. Also: target, host site, spam, backdoor, exploit, explore, sell, launder, pay %, morph

Alternative approach

Embrace the fiction

What problems are you trying to solve?

What fictional definitions of personhood will be useful?

- needn’t be conscious
- e.g. yet-to-be-conceived children, linghams, rivers

Check for unwanted side-effects

Legal Status Criteria for AI: Let's not give rights to malware

Miranda Mowbray
University of Bristol
miranda.mowbray@bristol.ac.uk