Human Language Technology for Active Defense against Social Engineering

Bonnie J. Dorr

with special acknowledgment to: Tomek Strzalkowski, Samira Shaikh, Adam Dalton, Larry Bunch, Amir Masoumzadeh, Alan Zemel, Sashank Santhanam, Archna Bhatia, and Brodie Mather

ihmc

RPI, SUNY Albany, UNCC

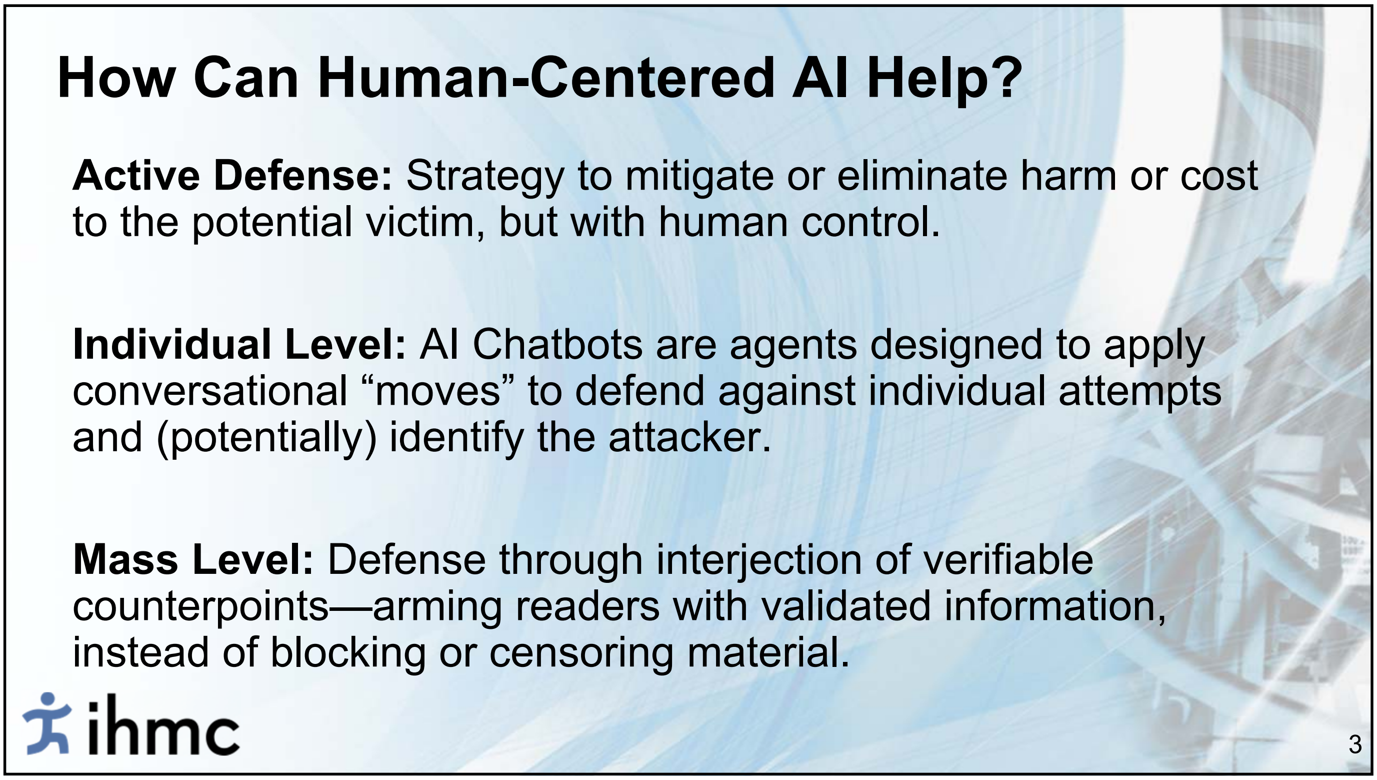
What Is Social Engineering?

Definition: Human interaction that manipulates people into divulging confidential information for access to systems, networks, locations, or for financial gain.

Individual Level: Use of deception to manipulate individuals into responding under pretexts, often divulging confidential or personal information that may be used for fraudulent purposes

Mass Level: Broad dissemination of scams or information with questionable validity, e.g. through viral social media posts.

ihmc

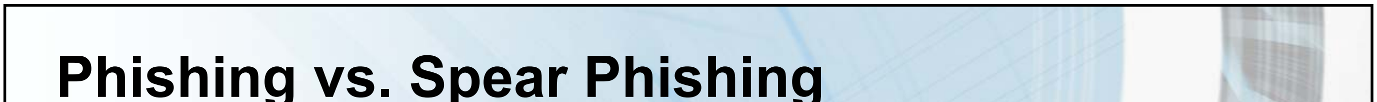


How Can Human-Centered AI Help?

Active Defense: Strategy to mitigate or eliminate harm or cost to the potential victim, but with human control.

Individual Level: AI Chatbots are agents designed to apply conversational "moves" to defend against individual attempts and (potentially) identify the attacker.

Mass Level: Defense through interjection of verifiable counterpoints—arming readers with validated information, instead of blocking or censoring material.



Phishing vs. Spear Phishing

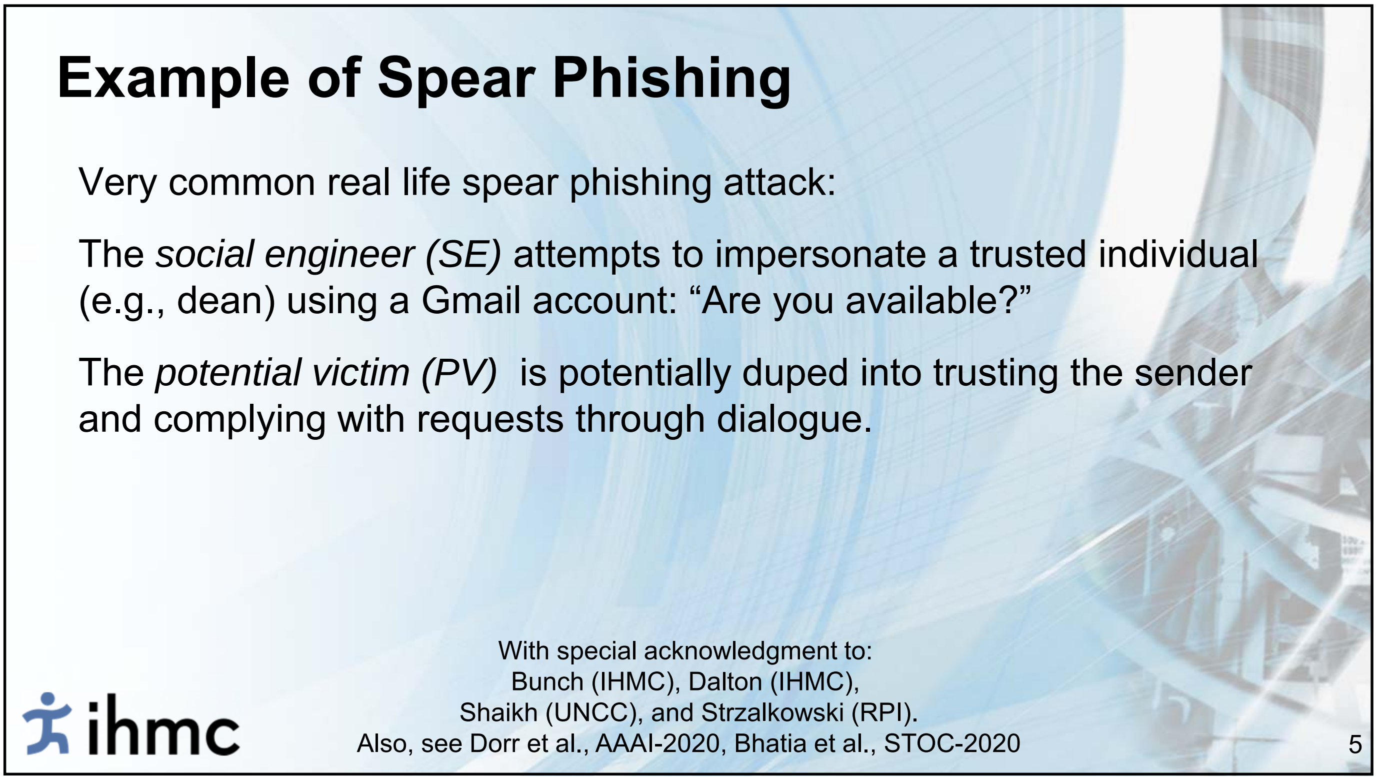
What is Phishing?

- A scam where a social engineer attempts to obtain private information, often via generalized communications.
- Low response rate: ~3% opened emails, 5% success rate of those opened.

What is Spear Phishing?

- A more focused phishing technique where SE poses as a trusted individual (usually a superior) and engages in highly customized communications to PVs.
- High response rate: ~70% opened emails, 50% success rate of those opened.

Acronym definitions:
SE = Social Engineer
PV = Potential Victim



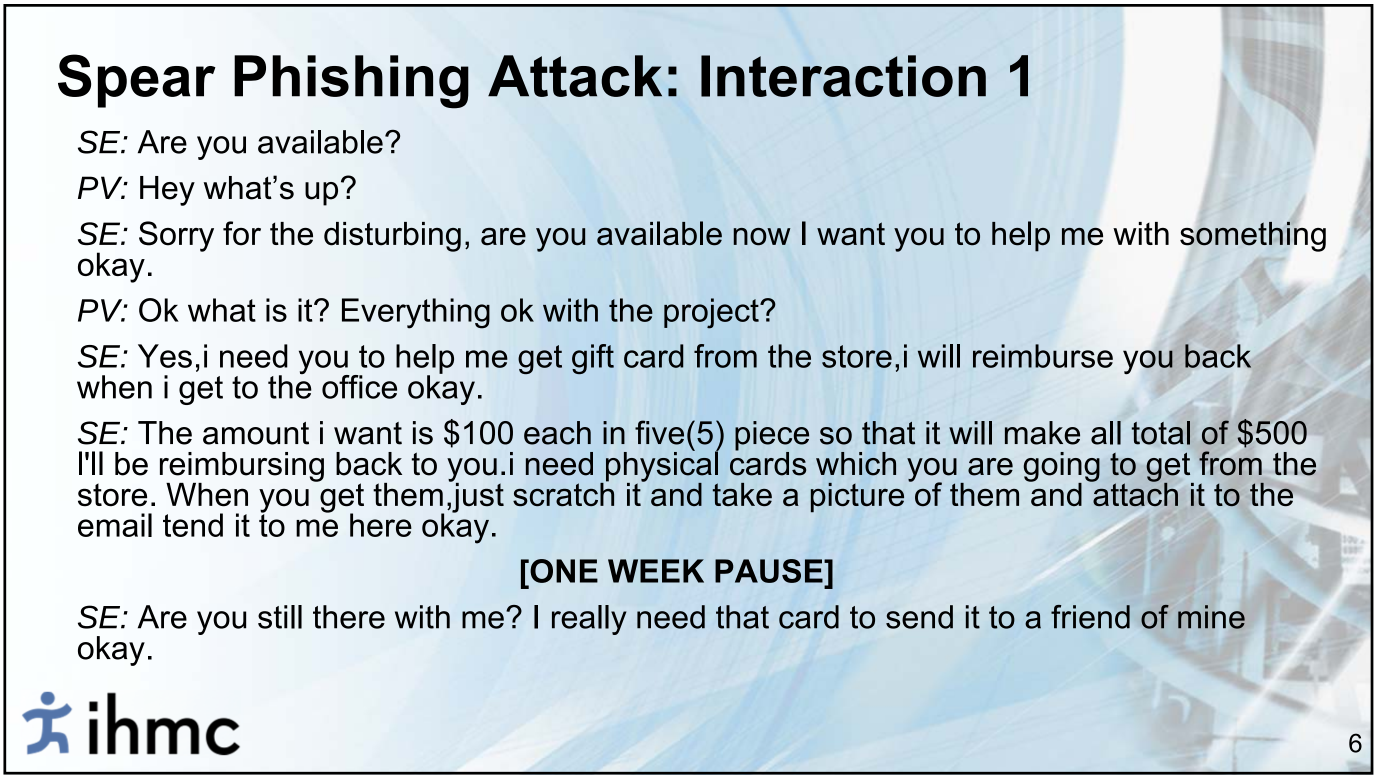
Example of Spear Phishing

Very common real life spear phishing attack:

The *social engineer (SE)* attempts to impersonate a trusted individual (e.g., dean) using a Gmail account: “Are you available?”

The *potential victim (PV)* is potentially duped into trusting the sender and complying with requests through dialogue.

With special acknowledgment to:
Bunch (IHMC), Dalton (IHMC),
Shaikh (UNCC), and Strzalkowski (RPI).
Also, see Dorr et al., AAAI-2020, Bhatia et al., STOC-2020



Spear Phishing Attack: Interaction 1

SE: Are you available?

PV: Hey what’s up?

SE: Sorry for the disturbing, are you available now I want you to help me with something okay.

PV: Ok what is it? Everything ok with the project?

SE: Yes,i need you to help me get gift card from the store,i will reimburse you back when i get to the office okay.

SE: The amount i want is $100 each in five(5) piece so that it will make all total of $500 I'll be reimbursing back to you.i need physical cards which you are going to get from the store. When you get them,just scratch it and take a picture of them and attach it to the email tend it to me here okay.

[ONE WEEK PAUSE]

SE: Are you still there with me? I really need that card to send it to a friend of mine okay.

Spear Phishing Attack: Interaction 2

PV: Hey still want these cards?

SE: Yes have you get it?

PV: Sure, where to send them?

SE: send them here for me attach it with massage here okay.

SE: just scratch it and take a picture of them and attach it to the email tend it to me here okay.

PV: You said you needed physical cards – need your mailing address to send

SE: send it for me here now okay.

PV: And where is here? Your home address?

SE: i mean here not my house just take pictures of each of them and send it okay.

ihmc

Spear Phishing Attack: Interaction 3

PV: Uh sorry eventually got to it – is this what you want? Can you clarify?

VIP
CARD
Club Logo
MEMBERS ONLY
MEMBER NAME
000-000-000
MEMBER SINCE
01/01
VILID THRU
01/01

SE: yes do you get send it for me okay.

SE: open the envelope and scratch it and take the picture of every each and send it for me here now okay. I am waiting for make it **fast** as you can please.

ihmc

Spear Phishing Attack: Interaction 4

SE: Hello have you get it send me the card I am waiting for it since make it **fast** okay.

PV: Yes sorry just worried if it's safe to send these cards via email

SE: yes it's okay with me do it for I have been waiting do I **fast** okay.

PV: Here you go

Lorem ipsum
Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat. Ut wisi enim ad minim veniam, quis nostrud exerci tation ullamcorper suscipit lobortis nisl ut aliquip ex ea commodo consequat. Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero eros et accumsan et iusto odio dignissim qui blandit praesent luptatum zzril delenit augue duis dolore te feugait nulla facilisi. Nam liber tempor cum soluta nobis eleifend option congue nihil imperdiet doming id quod mazim placerat facer
Club Logo
Address
Lorem Ipsum Building
111 Avenue, 1st Floor
Dolor Sit, LM 00000
1-111-111-1111
name@domain.com
Club Logo
Club Logo
Club Logo

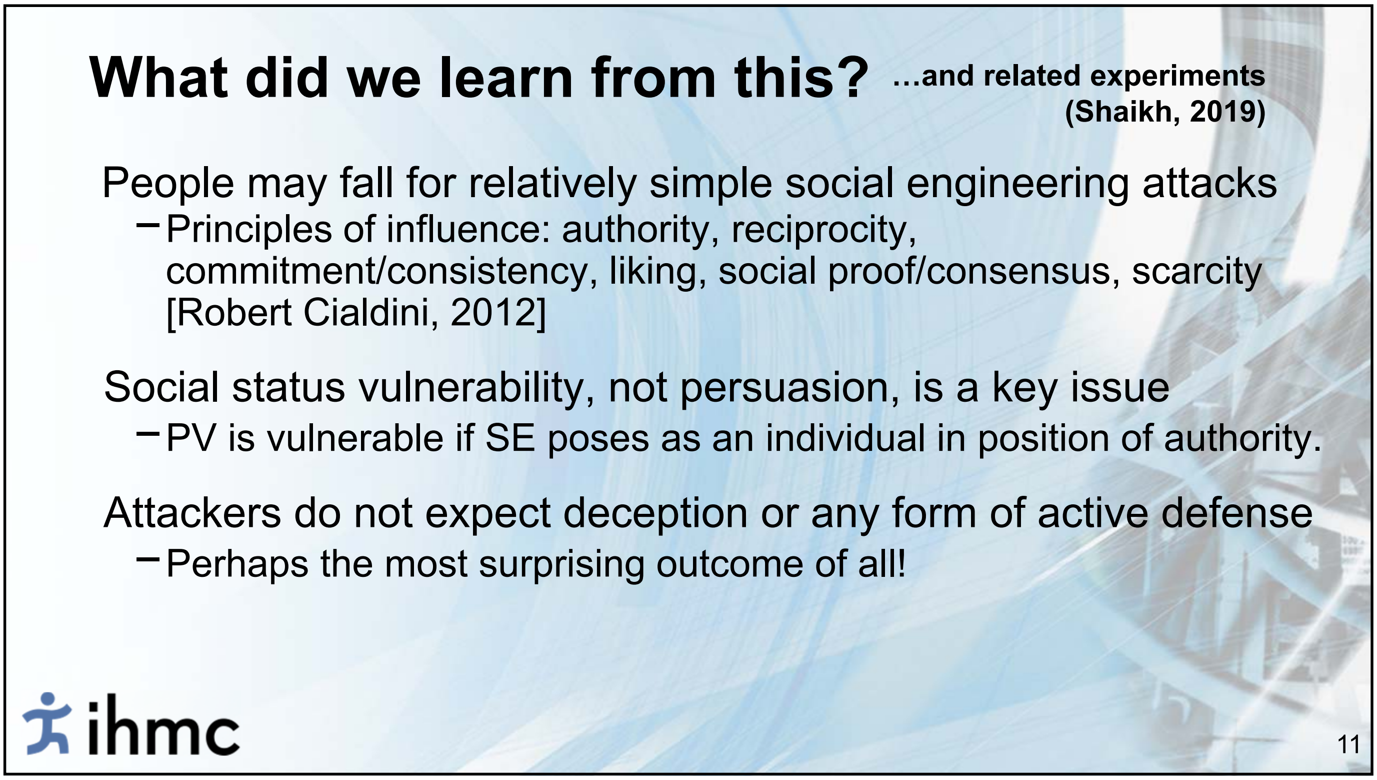
SE: I get the card but it is not clear you don't scratch it very well scratch it well and resend it and I told you to get me five pieces of the card you only send me one and is not clear do it **fast** as you can okay I am waiting for it.

ihmc

Summary of Information Gathered by Active Defense

Time Zone:	Central Europe, West Africa
Device:	Android 4.4.4 Phone (circa 2014)
	ARM v7l (32 bit), 1GB
Language:	US English
Connection:	Cellular 3G/4G

ihmc



What did we learn from this? **…and related experiments (Shaikh, 2019)**

People may fall for relatively simple social engineering attacks

- Principles of influence: authority, reciprocity, commitment/consistency, liking, social proof/consensus, scarcity [Robert Cialdini, 2012]

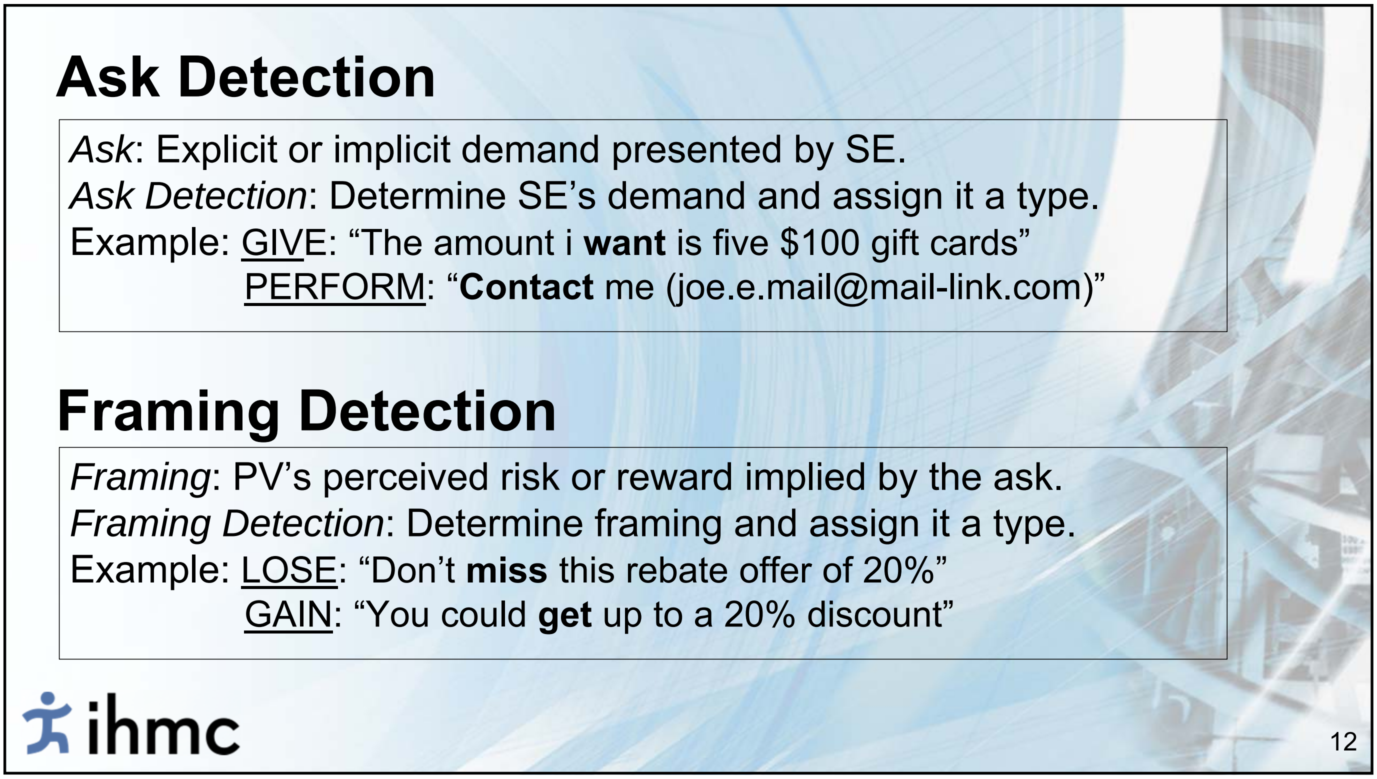
Social status vulnerability, not persuasion, is a key issue

- PV is vulnerable if SE poses as an individual in position of authority.

Attackers do not expect deception or any form of active defense

- Perhaps the most surprising outcome of all!

ihmc



Ask Detection

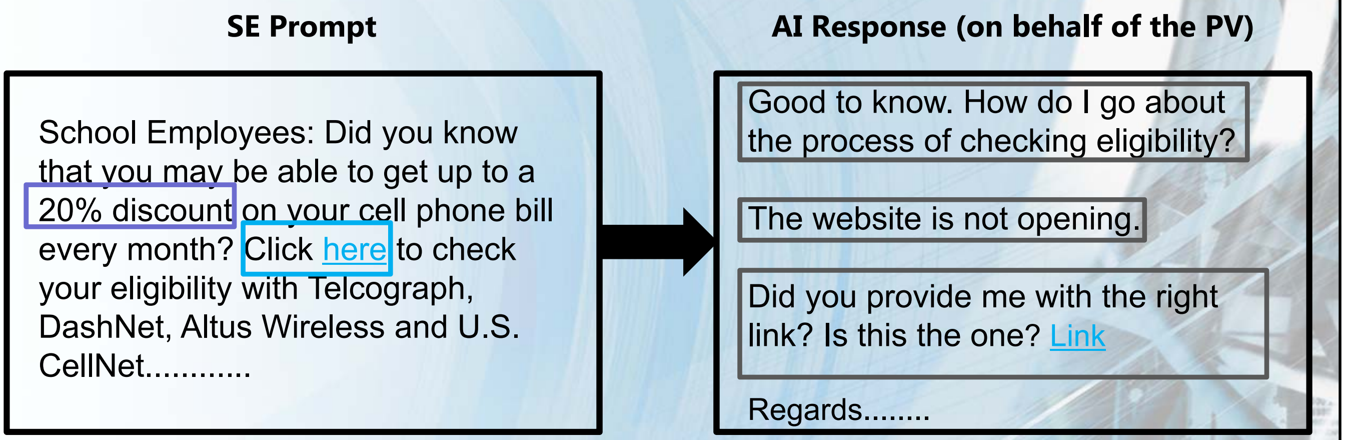
Ask: Explicit or implicit demand presented by SE.
Ask Detection: Determine SE’s demand and assign it a type.
Example: GIVE: “The amount i **want** is five $100 gift cards”
PERFORM: “**Contact** me (joe.e.mail@mail-link.com)”

Framing Detection

Framing: PV’s perceived risk or reward implied by the ask.
Framing Detection: Determine framing and assign it a type.
Example: LOSE: “Don’t **miss** this rebate offer of 20%”
GAIN: “You could **get** up to a 20% discount”

ihmc

AI-induced Response to Social Engineering Attack



SE Prompt

School Employees: Did you know that you may be able to get up to a 20% discount on your cell phone bill every month? Click here to check your eligibility with Telcograph, DashNet, Altus Wireless and U.S. CellNet............

AI Response (on behalf of the PV)

Good to know. How do I go about the process of checking eligibility?

The website is not opening.

Did you provide me with the right link? Is this the one? Link

Regards........

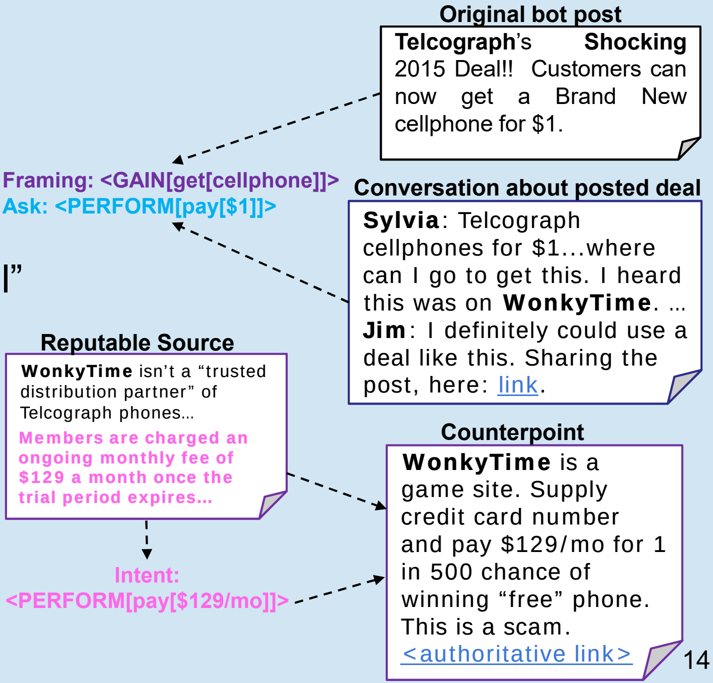
Framing: <GAIN[get[20%discount['finance_money']]]>

Ask: <PERFORM[check[eligibility(here)]]]

ihmc

Social Engineering at scale: mass propagation of scams

- Generalization and adaptation of individual-level technologies to mass scale applications
- Social Media: Bots employed to pump out posts that later "go viral"
- Exploitation of AI technology to position active defenses and slow, prevent, and counter the social engineering impact



Original bot post

Telcograph's **Shocking** 2015 Deal!! Customers can now get a Brand New cellphone for $1.

Framing: <GAIN[get[cellphone]]>
Ask: <PERFORM[pay[$1]]>

Conversation about posted deal

Sylvia: Telcograph cellphones for $1...where can I go to get this. I heard this was on **WonkyTime**. ... **Jim**: I definitely could use a deal like this. Sharing the post, here: link.

Reputable Source

WonkyTime isn't a "trusted distribution partner" of Telcograph phones...

Members are charged an ongoing monthly fee of $129 a month once the trial period expires...

Intent:
<PERFORM[pay[$129/mo]]>

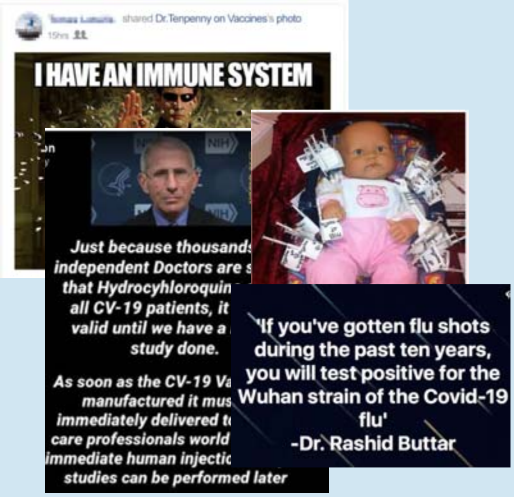
Counterpoint

WonkyTime is a game site. Supply credit card number and pay $129/mo for 1 in 500 chance of winning "free" phone. This is a scam. <authoritative link>

ihmc

What about misinformation campaigns?

Original bot posts and memes



Conversations about posts and memes

Jan: Hydroxy ruled out cuz there is no major study, but as soon as C19 vax is available, they will immediately inject? **Sue**: Hydroxy a better solution. **Tim**: Everyone is too scared to go out til they get a vaccine, and it doesn't have to work! **Mary:** Agree! Dr. F. pure evil. Gets his cut of vaccines. **Emma**: Vaccines contain antifreeze! **Harry**: No thanks, I'll take my chances. **Bill**: Yup. And I heard flu vax increases C19 risk. **Jean:** Fight mandatory vax!

Reputable Sources

FDA Study: Hydroxy chloroquine associated severe risks, including

US Armed Forces Study no evidence of interference flu vaccine with C19

CDC study propylene g safe for medications

NIH-Funded Research in Am Journ Public Health: **Foreign Trolls Amplify Vaccine Debate**

Intent: <PERFORM[amplify[vaccine debate]]>

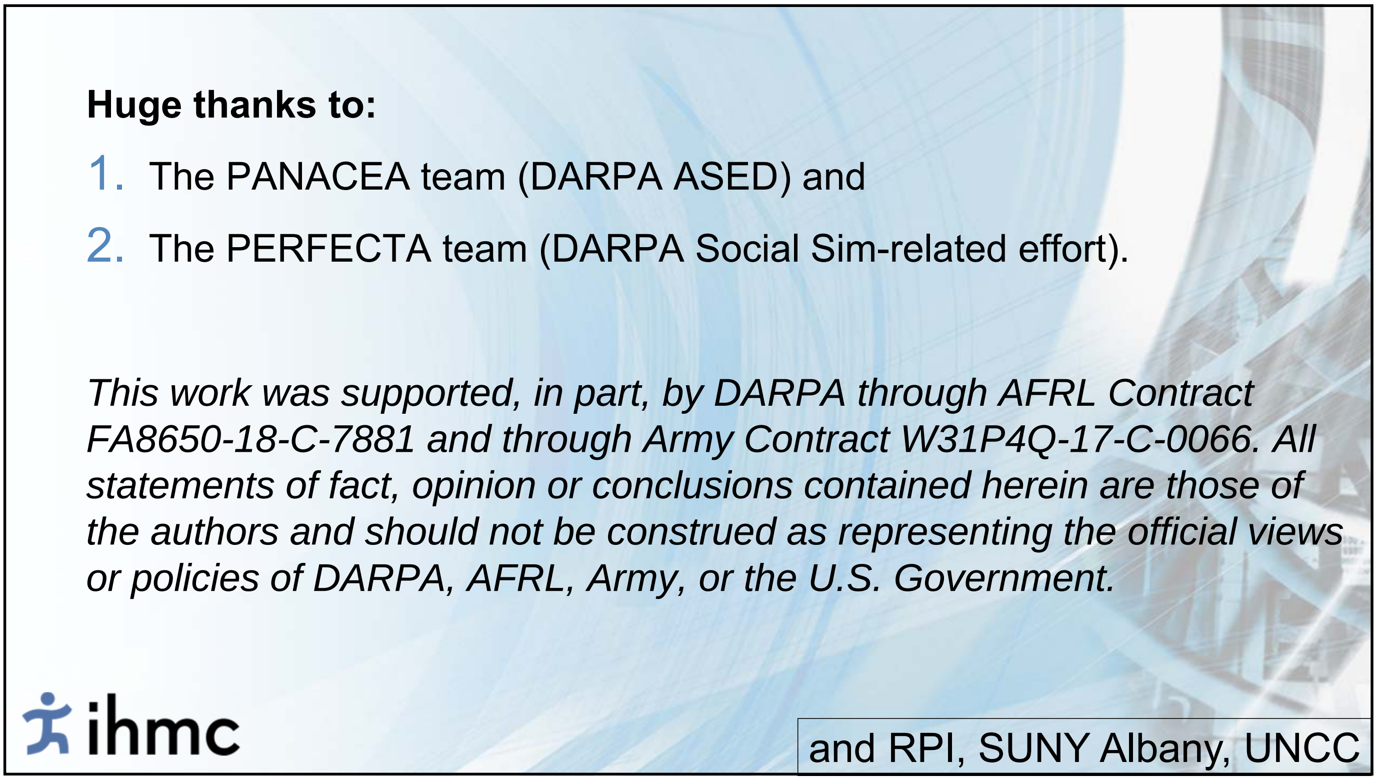
Counterpoint

FDA warns against use of hydroxy chloroquine for COVID-19 patients due to severe risks including death. Vaccines are not rushed for approval. Proper FDA protocol is followed. Polyethylene glycol is used, not antifreeze, deemed safe by CDC. US armed forces study refers to coronavirus but shows no evidence of connection between influenza vaccines and C19. NIH-funded study reveals sophisticated bots of foreign origin propagate misinformation to amplify vaccine debate. <authoritative links>

Framing: <LOSE[have[free choice]]>
Framing: <GAIN[pose[danger to health]]>
Ask: <PERFORM[fight[mandatory vax]]>

Human-Centered AI as Alternative to Censorship for Social Engineering

- Insert counter-point before too much damage has been done.
  - In the control of the user, able to approve or edit AI-generated responses
- Is censorship an option?
  - Remove or block misinformation, e.g., hide/ban channels (e.g., anti-vax posts)
  - Is this socially responsible?
    - PRO: Some have argued that these social media cannot be "vehicles for misinformation".
    - CON: Some have argued that misinformation is a bigger problem than censorship can solve.
- Alternatives to censorship?
  - Tackle misinformation with information and let the human decide
  - Human-centered AI solution to social engineering on mass scale
    - Detect misinformation automatically
    - Provide human author a counter-point from verifiable, trusted sources
    - The human reader gets to decide!



Huge thanks to:

1. The PANACEA team (DARPA ASED) and
2. The PERFECTA team (DARPA Social Sim-related effort).

This work was supported, in part, by DARPA through AFRL Contract FA8650-18-C-7881 and through Army Contract W31P4Q-17-C-0066. All statements of fact, opinion or conclusions contained herein are those of the authors and should not be construed as representing the official views or policies of DARPA, AFRL, Army, or the U.S. Government.

ihmc

and RPI, SUNY Albany, UNCC