APPG
Cyber Security - Online Harms

Dr Konstantinos Mersinas

What is harmful *content* and *activity* online?

Harmful content = Content crime

Not all harmful content is illegal

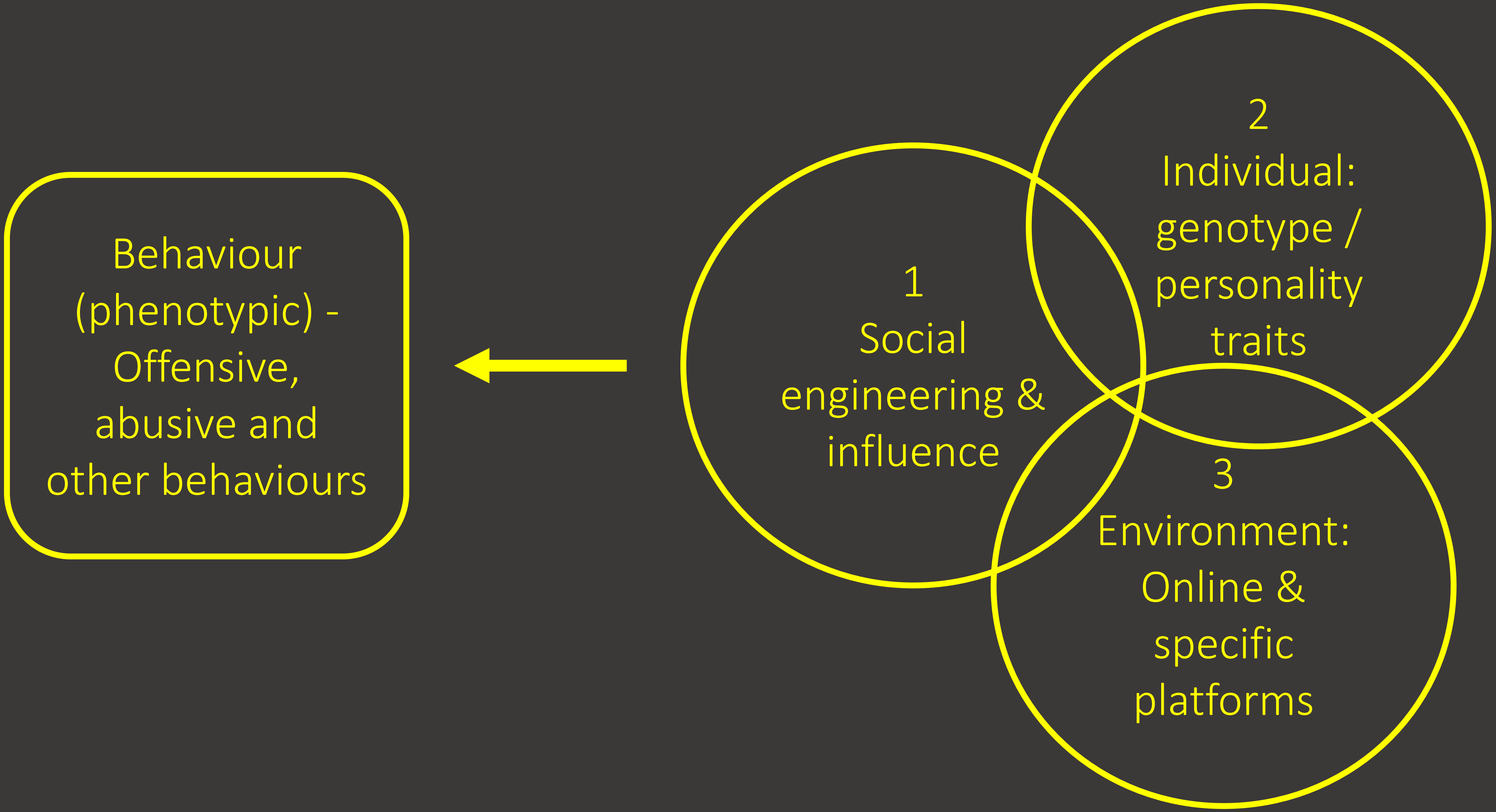
- Obscene & indecent content (child sexual abuse materials)
- Live distant child abuse
- Extreme & revenge pornography
- Selling stolen personal data (eWhoring)
- Hate speech
- Misinformation & disinformation

Harmful activity = Interpersonal offenses

- Cyberbullying
- Cyberstalking
- Grooming
- Child sexual exploitation
- Sexual coercion and extortion

- Vulnerable groups: children, adolescents, the elderly
(Hub for Intergenerational Vulnerability to Exploitation - HIVE)

Explaining online harms: 3 Components



Behaviour (phenotypic) - Offensive, abusive and other behaviours

1 Social engineering & influence

2 Individual: genotype / personality traits

3 Environment: Online & specific platforms

Component 1: Social engineering & influence

Component 1: Social engineering & influence

The use of deception to manipulate individuals into revealing confidential or personal information that may be used for fraudulent purposes

- Impersonation & deception

Component 1: Social engineering psychological principles

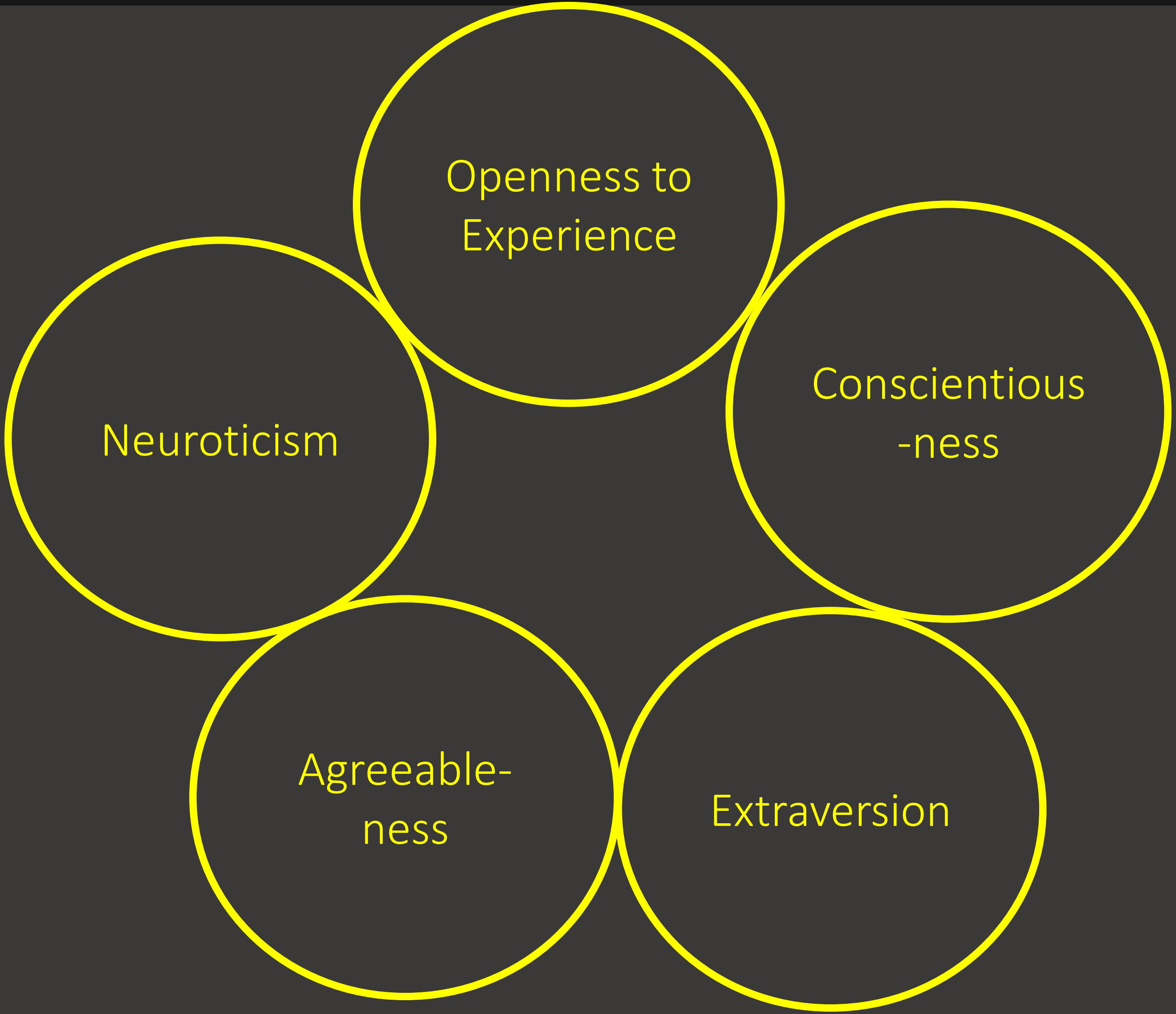
1.	**Influence**	(communication skills, language, frequency, listening)
2.	**Reciprocation**	(targeted 'gifts')
3.	**Commitment** & **Consistency**	(statements in public)
4.	**Social Proof**	(peer pressure, the group)
5.	**Liking**	(attractiveness, common interests, ideas)
6.	**Authority**	(less apparent online, except in groups)
7.	**Scarcity**	(actions presented as opportunities, time, resources)

Component 1: Social engineering & influence examples

- Underaged users distributing self-generated materials / on-camera actions
- Grooming: creating materials & meetings
- Peer pressure & commitment to a cause/action: radicalisation, hate speech
- Behaviour in groups / ‘packs’
- Disinformation (e.g. based on false authority)
- COVID-19 vulnerabilities: information seeking, working from home

Component 2: Individual & personality traits

Component 2: OCEAN Personality traits



Openness to Experience

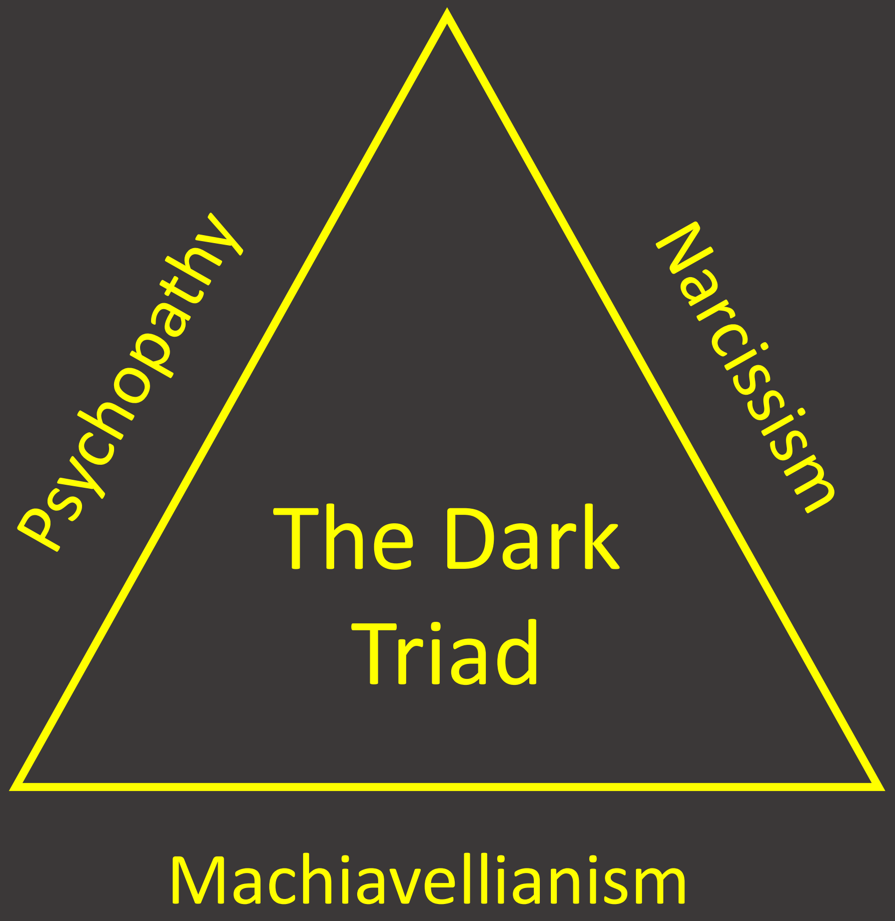
Conscientious-ness

Extraversion

Agreeable-ness

Neuroticism

Component 2: Personality traits



- **Narcissism**:

  feelings of superiority and entitlement

- Machiavellianism:

  manipulating others, concealed aggression

- Psychopathy:

  antisocial, lack of empathy, impulsivity

Component 2: Personality traits - identifiable, measurable behaviours

Cyberbullying / hate speech

Low: A, C
High: N, Dark Triad

Cyberstalking

High: E, O
(risk-seeking)

High: O, N
Low: E (inability to delay gratification)

Grooming

High on 3 traits of the Dark Triad

Criminal and analogous activities

High: N
Low: C, A, E
(low self-control)

Deception

Correlated to the dark triad

High-stakes lies predicted by Machiavellianism

Cyberbullying victims

High: N, O

O: Openness to Experience
C: Conscientiousness
E: Extraversion
A: Agreeableness
N: Neuroticism

Component 3: Environment

Component 3: Environment - the online disinhibition effect

- Factors: [White paper: anonymous abuse]
  - dissociative anonymity
  - invisibility
  - asynchronicity
  - (minimisation of) authority

- Negative: offensive behaviour, hate speech
- Positive: free expression

Component 3: Environment & platform design

Behavioural functionality (boosts and nudges)

[White paper: Safety by design, mechanisms to allow users to report content]

EAST:

- Easy
- Attractive
- Social
- Timely

MAT:

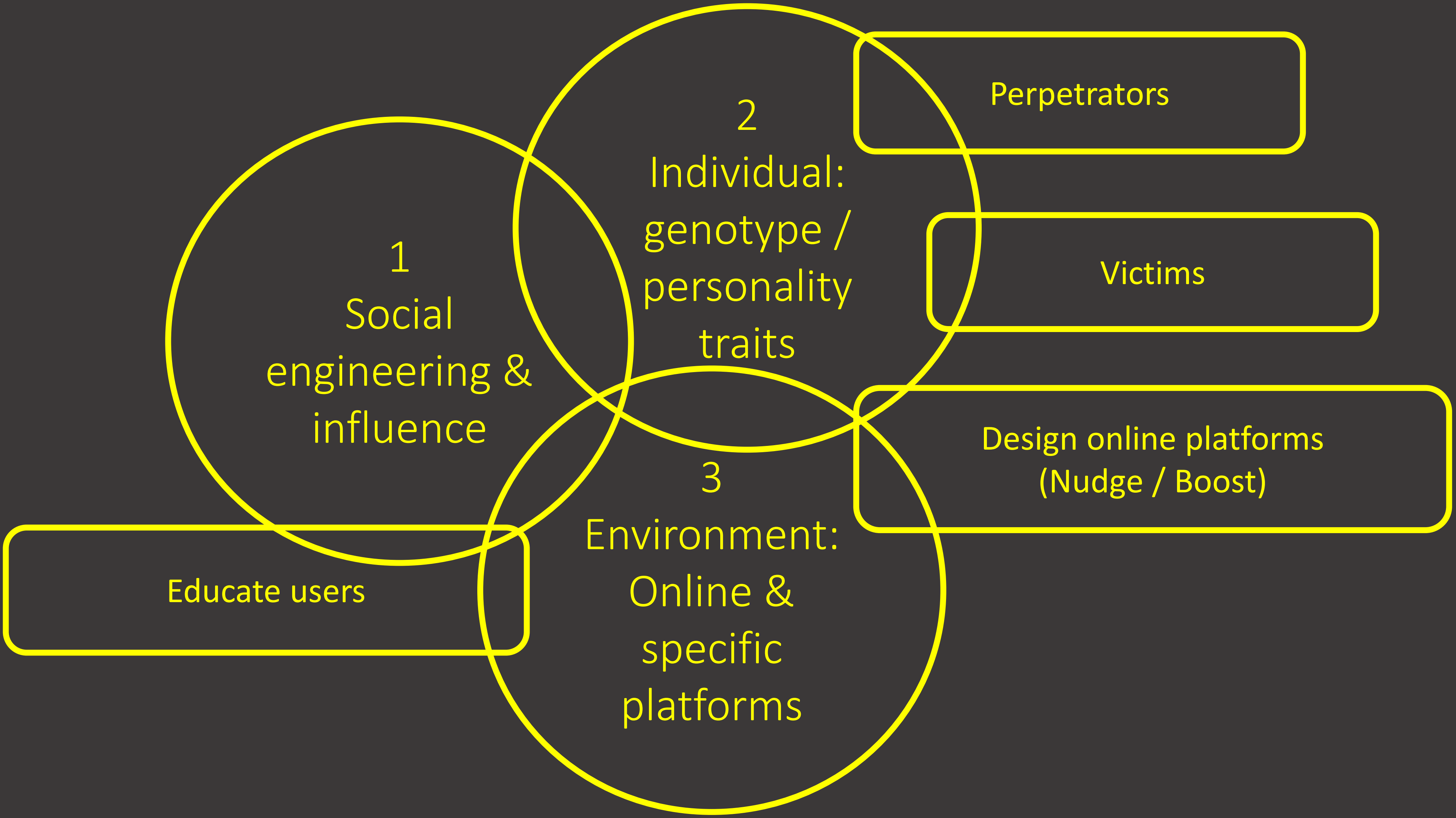
- Motivation
- Ability
- Trigger

Component 3: Environment: Platform design & AI

- AI not a panacea
- AI as a solution working *with* users
  - User behaviour change via an AI assistant

[White paper: What part will technology, education and awareness play in the solution?]

Solutions?



1
Social engineering & influence

2
Individual: genotype / personality traits

3
Environment: Online & specific platforms

Perpetrators

Victims

Design online platforms
(Nudge / Boost)

Educate users

Thank you!