Centre for Digital Trust and Security

Seedcorn Interim Progress Report 22/23

1. Project Title: **Rebuilding Democratic Discourse: Online Harms and Trust**

Project Investigators:

Mihaela Popa-Wyatt (PI)
Justina Berskyte (CO-I)
Graham Stevens (CO-I)

Project overview:

The overall project, of which this seed project is one part, has five objectives. These are anchored in philosophy of language, social and political philosophy, informed by linguistics, computer science and social sciences. The five objectives are: 1) to create a taxonomy of harmful speech; 2) to conduct case studies of specific types of harmful speech, analysing the available corpora; 3) modelling how harmful speech spreads; 4) devising countermeasures; 5) drawing where the boundaries of countermeasures and regulation ought to lie. In this seed grant, we focus on objective 2. Specifically, on one case study: hate speech in the incel community.

Key milestones to date:

We have articulated three philosophical theoretical hypotheses - that the language used in these communities serves three functions:

1. Incels use derogatory language towards women ('Stacy', 'femoid') to express hostility toward seemingly unattainable women. Incels also derogate other men, designating 'alpha' males as 'Chads'. This serves to enforce the group-boundaries (us vs them). The names applied by incels allocate social roles during a discourse which support the incel belief system.
2. Incels use language that signals group-identity and reinforces their status as an 'incel'. This cements group-membership and loyalty to incel-ideology. Often, they weave a narrative of victimhood where unattainable women and attractive men are perceived as the 'enemy' and 'rival' (respectively), while

incels are the 'victims' for not receiving social and emotional goods they see themselves entitled to.

3. Incels use language that polices group-boundaries, drawing a wedge between 'truecels' and 'fakecels', to deter and outcast those who do not fully embrace incel-ideology. We will explore how such policing leads to tribalism and conspiratorial thinking, which is reinforced by echo-chambers.

Outputs to date:

Talks (specific on the language of "Incels")

- TBC. (Mihaela Popa-Wyatt & Justina Berškytė). The Digital Trust and Security Seminar Series. April 2024
- "Who are the incels: interpellation and persona?" (Justina Berškytė). Justice and Social Media Workshop. University of Southhampton. 15-16 Sep 2023.
- "Incel Persona: The Fallen Angel". (Mihaela Popa-Wyatt & Justina Berškytė). *Words Workshop*. University of Pittsburg. 31 July 2023.
- "Incel Ideology, Persona, and In-Between Identities". (Mihaela Popa-Wyatt &
- Justina Berškytė, Maurits Bekkers). Harmful Content (offline and online): Challenges and Ways Out. University of Manchester. 10-11 July 2023.
- "Messaging the UnderDog" (Mihaela Popa-Wyatt & Justina Berškytė), Workshop on Themes in Oppressive Language, University of Manchester. 3 March 2023.
- "How Incel Ideology Becomes a Cult" (Justina Berškytė & Mihaela Popa-Wyatt). STAL. Université de Sorbonne. 7-9 June 2023. (accepted, could not travel)
- Poster of the project, displayed at the CDTS event, 6 July 2023.
- "Incels: Identity and Group-Boundary Policing" (Justina Berškytė & Mihaela Popa-Wyatt). *Aristotelian Society. Joint Session*. The University of London. 7-9 July 2023. (submitted, not accepted)

General Talks (on the broader objectives related to online harms)

- Mitigating unfairness. (Mihaela Popa-Wyatt). *Complexity Theory, Social Ontology, and Social Change Workshop*, MIT. 6-7 October 2023.
- The contagion of online harms. (Mihaela Popa-Wyatt). Justice and Social Media Workshop. The University of Southhampton. 15-16 Sep 2023.

- What role does authority vs power play in oppression? (Mihaela Popa-Wyatt). *Workshop on Authority, Power, and Accountability*. The University of Stockholm. 15 Aug 2023.
- "How Hate Speech Works" (Mihaela Popa-Wyat). Royal Institute of Philosophy. 10 February 2023. (YouTube video https://www.royalinstitutephilosophy.org/event/how-hate-speech-works/)
- "Slurs and Mocking Laughter: a unified account via Interactional History" (Jonathan Ginzburg & Mihaela Popa-Wyatt). STAL. Université de Sorbonne. 7-9 June 2023.
- "Online Hate. Is Hate an Infectious Disease? Is Social Media a Promoter?" (Mihaela Popa-Wyatt), Workshop "Political Normativity and Ethics". The University of Granada. 9-12 May 2023.
- "The Spread of Online Hate: Complex Contagion" (Mihaela Popa-Wyatt). Pacific APA, San Francisco, 5-8 April 2023.
- "Hate Speech and Free Speech". PPE Seminar. The University of Manchester. 7 December 2022.
- "How Norms Shift". The University of Durham. 8 February 2023.
- "Oppressive Speech (Constitutive vs Causal)". The University of Manchester. Politics Department. 23 November 2022.
- "A hierarchical game-theoretic model of social oppression by slurring". The University of Manchester. Philosophy Department. 16 November 2022.
- "Games, Oppression and Oppressive Speech". The University of Washington. Philosophy Department. 7 November 2022.
- "Spreading mechanisms of harmful language on the net: Anatomy of two shitstorms". (Tatjana Scheffler, Veronika Solopova & Mihaela Popa-Wyatt) "Verbreitungsmechanismen schädigender Sprache im Netz: Anatomie zweier Shitstorms". University of Ruhr-Bochum. 1 December 2022.

Policy brief:

- "The challenges of regulating online speech". Policy@Manchester 27 July 2023 https://blog.policy.manchester.ac.uk/posts/2023/07/the-challenges-of-regulating-online-speech/
- Submission of a policy pitch video for MCSA policy-award (Mihaela Popa-Wyatt)

Peer-Reviewed Papers (general on oppression and online harms)**:**

- "Online Hate. Is Hate an Infectious Disease? Is Social Media a Promoter?" (Mihaela Popa-Wyatt). *Journal of Applied Philosophy*. 2023. DOI:10.1111/japp.12679
- "Norm-Shifting through Oppressive Acts". In *Mind, Language, and Social Hierarchy: constructing a shared social world*. Editors: Sally Haslanger, Karen Jones, Greg Restall, François Schroeter, and Laura Schroeter. Oxford University Press. Forthcoming 2024.

Invited blog contributions (to be completed):

- "Democratized Speech and the Spread of Social Ills". *The Philosopher's Magazine*. (Mihaela Popa-Wyatt)
- The JAP - a collaboration with the blog Justice Everywhere - publish short blog posts on papers published in the journal.
- Blog Series for the APA focused on Philosophy and Technology
- Pitch on the JAP paper for New Work in Philosophy, a Substack page edited by Barry Maguire and Marcus Arvan.

Submitted Public Philosophy Blogs, Abstracts, Draft Papers:

"How to become an incel" (Maurits Bekkers). Blog submitted to *Open for Debate.* 10 August 2023 (awaiting decision)

"The Social Role of an Incel". (Justina Berškytė & Mihaela Popa-Wyatt) Abstract submitted for Interdisciplinary Workshop "Democratisation, roles, and discourse", 12 September 2023, Rabinstraße 8, University of Bonn.

"Responsibility for hate speech online". (Mihaela Popa-Wyatt) abstract submitted to 'Understanding Offence: Delimiting the (Un)sayable' - a multidisciplinary conference to be held at the Institute of Advanced Study. University of Durham. April 2024.

University of Lisbon. 2024.

"Who are the Incels? Interpellating the Incel Persona" (draft paper to be submitted to *Journal of Ethics and Social Philosophy*) (Justina Berškytė & Mihaela Popa-Wyatt)

We aim to pitch an article on incels or online communities for a public philosophy Op-ed *Aeon* (edited by Nigel Warburton). (Justina Berškytė & Mihaela Popa-Wyatt)

Abstract submitted for a Special Issue on Taboo. *Languages*. Edited by Keith Allan & Andrea Pizarro. (Justina Berškytė & Mihaela Popa-Wyatt, Adelina-Dalia Valoschi)

Popularise the empirical findings on incels. Monkey Cage. Washington Post.

Computational Agent Responsibility. University of Manchester. 20-22 September.

Analysis of social hierarchy and ranking within incel community. (collaboration with Veronika Solopova, Freie University).

Paper (work-in-progress) focuses on defining the characteristics of an incel persona, and how that is shaped by the misogynistic ideology that incel users subscribe to. (Mihaela Popa-Wyatt and Justina Berškytė).

Grants

- *New Investigator: Faculty Research Investment PI Fund* (£194.956,31): "Rebuilding Trust in Public Discourse: The Good Speech Project". The University of Manchester. (Mihaela Popa-Wyatt, PI), starting August 2023-2025
- Small Networking Grant (£1,978.50) (Mihaela Popa-Wyatt) - to disseminate results of the CDTS project and network with philosophers and political scientists at UCL in view of collaboration on future grants.

Workshop Organisation

Harmful Content (offline and online): Challenges and Ways Out. University of Manchester. 10-11 July 2023.
Organisers: Mihaela Popa-Wyatt, Justina Berškytė.
Invited Speakers: Jeffrey Howard, Emily McTernan, Mihaela Popa-Wyatt, Justina Berškytė, Maurits Bekkers, Alexander Brown, Jamie Mayerfeld, Lucy McDonald, Chris Cousens, Maxime Lepoutre.
Funding: SoSS Small Networking Grant (1,978.50), Centre for Digital Trust and Security.

Themes in Oppressive Language (Justina Berškytė & Mihaela Popa-Wyatt), University of Manchester. 3 March 2023.

Networking activities:

- Planned consultation for the EPSRC-funded project 'Digital Identity and Life-Course Study (DIALCS)', led by Yang Lu (York University) and David Buil-Gil (Manchester University).
- Liaised with Susan Benesch (the Berkman Centre, Harvard University) regarding possible collaboration with similar projects as part of “Rebooting Social Media”. (January 2023)
- Liaised with Jeffrey Howard, (Politics, UCL) regarding possible collaboration as part of his UKRI project on the Ethics of Digital Platforms (February 2023)
- Liased with Kesa White and Jennifer West regarding possible collaboration related to their work on extremist movements and incel communities, part of their project Polarization and Extremism Research and Innovation Lab (PERIL), American University (February 2023).
- Liaised with Jonathan Leader Maynard about the role of ideology in dangerous speech.
- Liaised with Suzanne Booth, at Policy@Manchester, to ask for guidance for writing a policy brief about interventions that may be relevant to the Online Safety Bill, and potentially connect with stakeholders.
- Liaised with Allysa Czerwinsky, Criminology, Manchester.
- Liaised with Lisa Sugiura, Sociology, Essex - discussion on countering misogyny online.
- Liaised with Shannon Valor (Ethics of Data and Artificial Intelligence), the Edinburgh Futures Institute.
- Liaised with Peter Coe, (Media Law), University of Birmingham.

Editorial (general themes related to harmful and misogynistic language)

- Mihaela Popa-Wyatt. (Under Contract/Reviewing stage). Harmful Speech and Contestation. Palgrave MacMillan. (Forthcoming 2024)
- Mihaela Popa-Wyatt. Oppressive Speech and Society (under review with Routledge).

UREC Research Ethics application - Approved in May 2023.

Outcomes achieved

The outcomes from the theoretical side are as expected. We have succeeded in building the ‘big picture’ of key social and psychological drivers of incel communities and ideologies. This large picture was clarified in various talks, blog (see above).

Our deliverables match the theoretical hypotheses we started with, and as researched by the politics and philosophy RAs. For example, we focus on how to

understand social capital as applied to internet/virtual communities, how to understand the function of the dehumanising language that incels use against women, and how this language shapes and is shaped by the ideology. Below we list the specific outcomes gained from the theoretical side.

Politics

· reviewed the literature on how gender-based derogatory language shifts social norms
· reviewed the literature on political and sociological features of incel-ideology
· identified social and epistemological mechanisms that lead to the formation and maintenance of group-identity and a tribal epistemology
· explored how group policing leads to tribalism and conspiratorial thinking, and how this is reinforced by echo-chambers

Philosophy

· reviewed the literature on how gender-based derogatory language normalises sexism and misogyny
· reviewed the literature on specific features of the harmful language in "incel" communities and relate this to philosophical theories of slurs and oppressive speech;
· articulated hypotheses about how the linguistic mechanisms identified produce social harms.
· identified mechanisms for signalling group-identity and group-boundary policing
· explained the social and epistemological mechanisms that lead to the formation and maintenance of group-identity and a tribal epistemology.

Papers writing preparation

Annotated bibliography (Maurits Bekkers and Adelina-Dalia Valoschi)
•Glossary of Derogatory language (incel.is)
•Reading list per topic
•Collection + summary of relevant articles (in shared dropbox)
•'Translation' of relevant political and sociological literature, to be applied to incel communities

***Deliverables from the empirical team**:*

CS team

· Develop automated techniques for corpus analysis to detect hateful (misogynistic) language;
· build classifiers for tagging/filtering data.

· use statistical analysis and sentiment analysis of the corpus produced by the Linguistics RA;
· use computational modelling and classifiers to tag data from the corpus in relation to a taxonomy of harmful language.
· compare the analysis of the language in “incel” to wider modelling of misogynistic language

Planned activities by project end:
We have secured external funds to organise an **Interdisciplinary Workshop** (Mihaela Popa-Wyatt, Justina Berškytė). June/July 2023. Report to be written in Sep 2023.

Activities

- Write up analysis and narrative to discuss the results from the corpus from the incel community, using statistical analysis of that corpus. (Lily Bromfield, Vera Hohaus, Leigh Harrington)
- analyse their speech using automated techniques for corpus analysis to detect hateful utterances. (Iqra Zahid, Riza Navarro)
- based on a classification of the acts performed with such speech, philosophers alongside linguists will work to explain the acts in terms of a theory of ideology and power, social roles and incel persona.

General future activities:

- Collaboration with Veronika Solopova (TU-Berlin) to write a piece on “banning” in incel communities.
- Collaboration with Juri Viehoff (Politics, University of Manchester) on issues of digital trust.
- Collaboration with Jeffrey Howard (Politics, UCL) on issues of free speech and digital harms.
- How The Light Gets In, Philosopher’s Zone.
- Liaise with charities, e.g. The Center for Countering Digital Hate, Stop Hate (UK), Some of US (San Francisco), Good Things.