Designing *Responsible* Natural Language Processing:

translating responsibility in AI through interdisciplinary reflection

Jacquie Rowe, Amanda Horzyk, Osman Batur İnce, Cyndie Demeocq

Supervisors: Benedetta Catanzariti, James Garforth, Hannah Rohde, Fabio Tollon

Designing Responsible Natural Language Processing

UKRI AI Center for Doctoral Training (CDT) in

Designing Responsible and Trustworthy Natural Language Processing (NLP) in-the-world

UK Research and Innovation

THE UNIVERSITY of EDINBURGH

Designing Responsible Natural Language Processing

Amanda Horzyk

Bríd-Áine Parnell

Cyndie Demeocq

Jacqueline Rowe

Jinzuomu Zhong

Kimberley Paradis

Neel Rajani

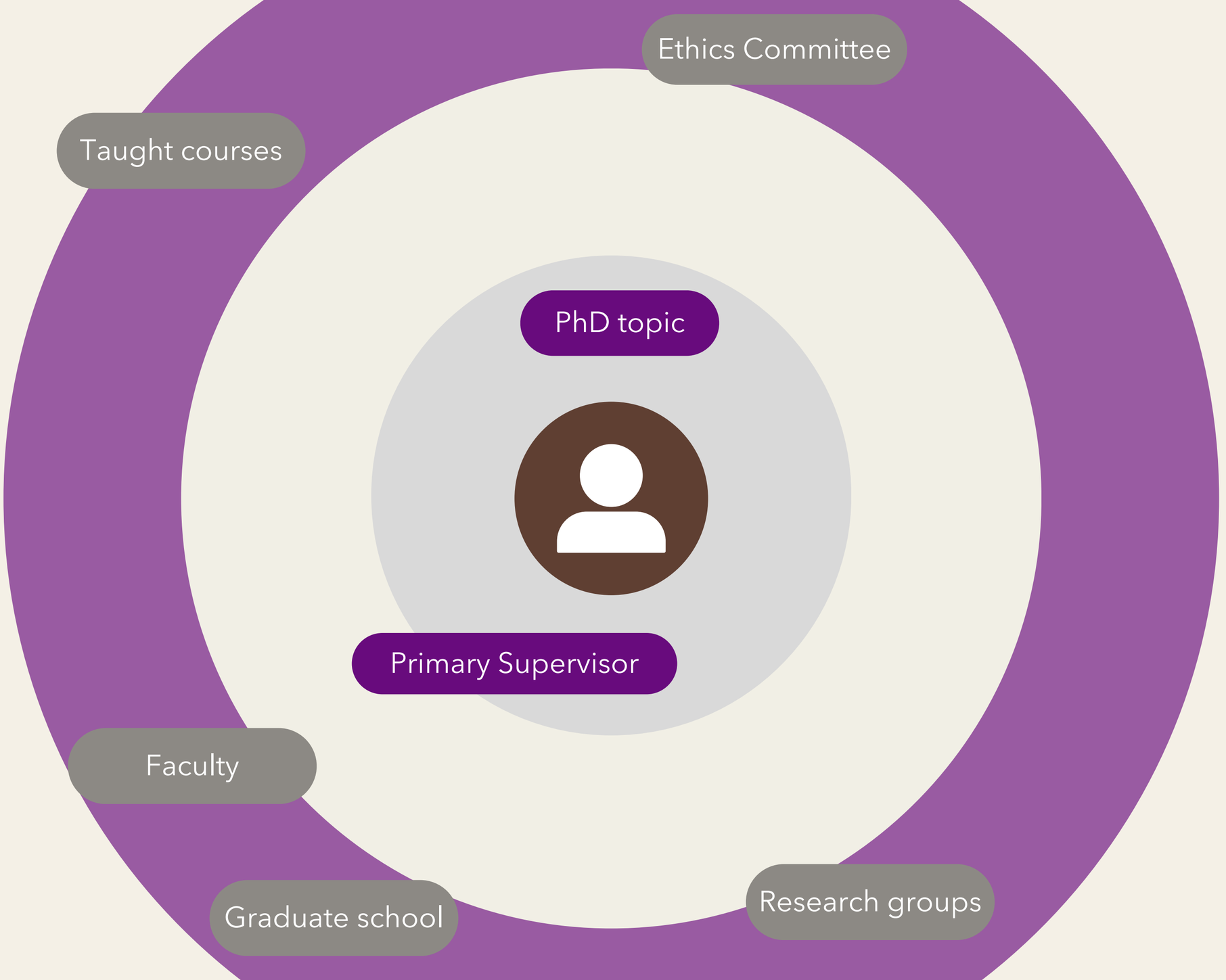
Osman Batur İnce

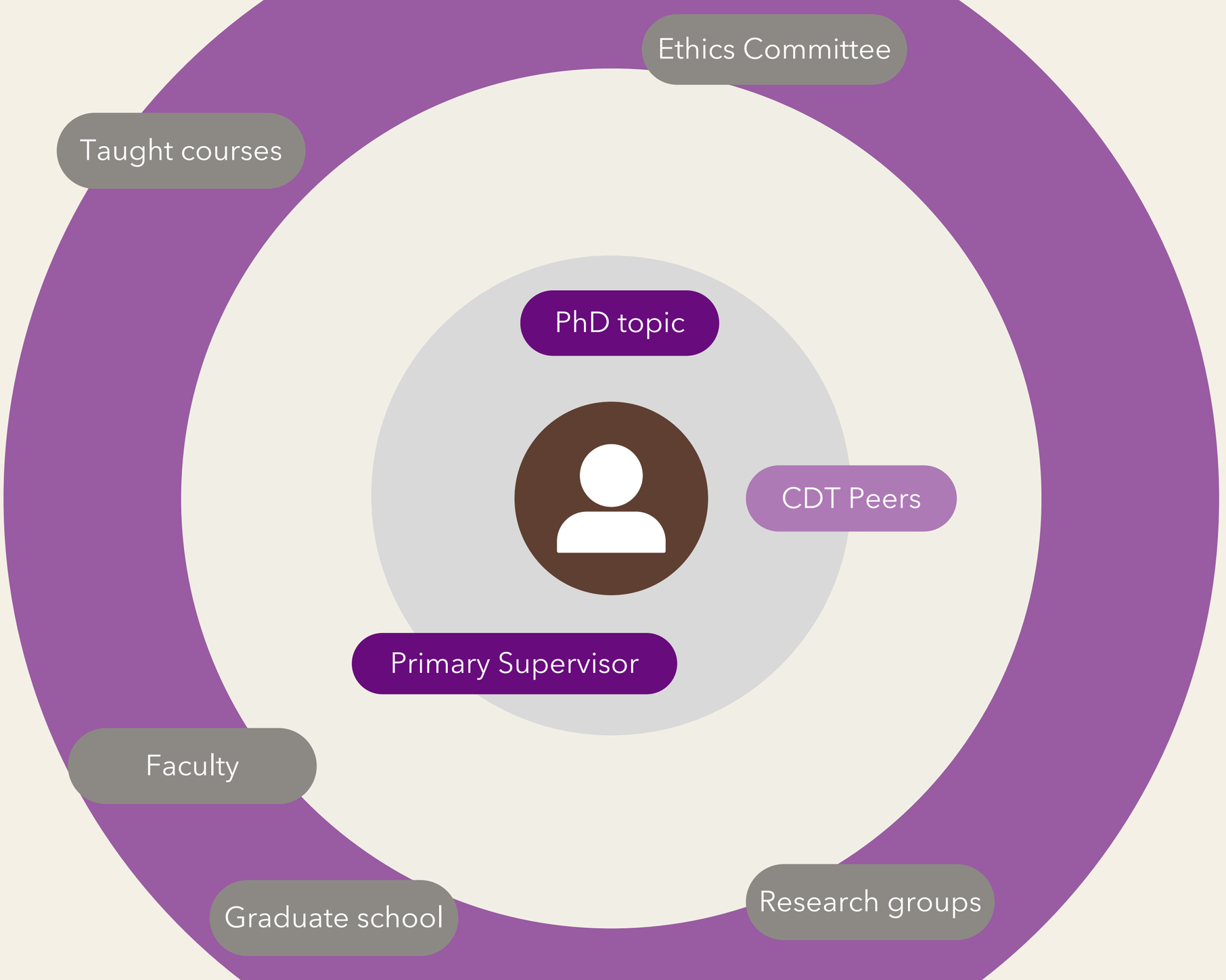
Rayo Verweij

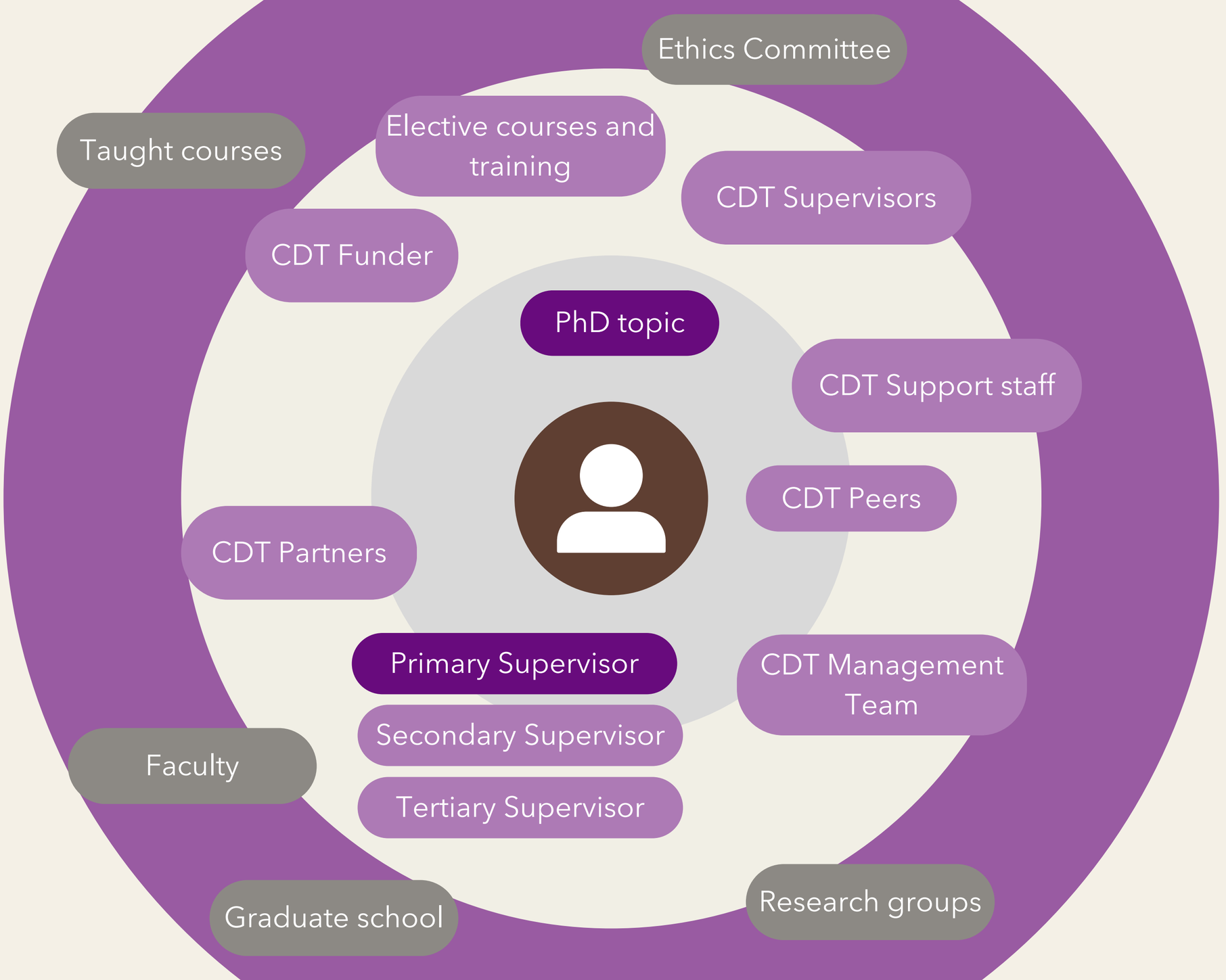
Sarah Immel

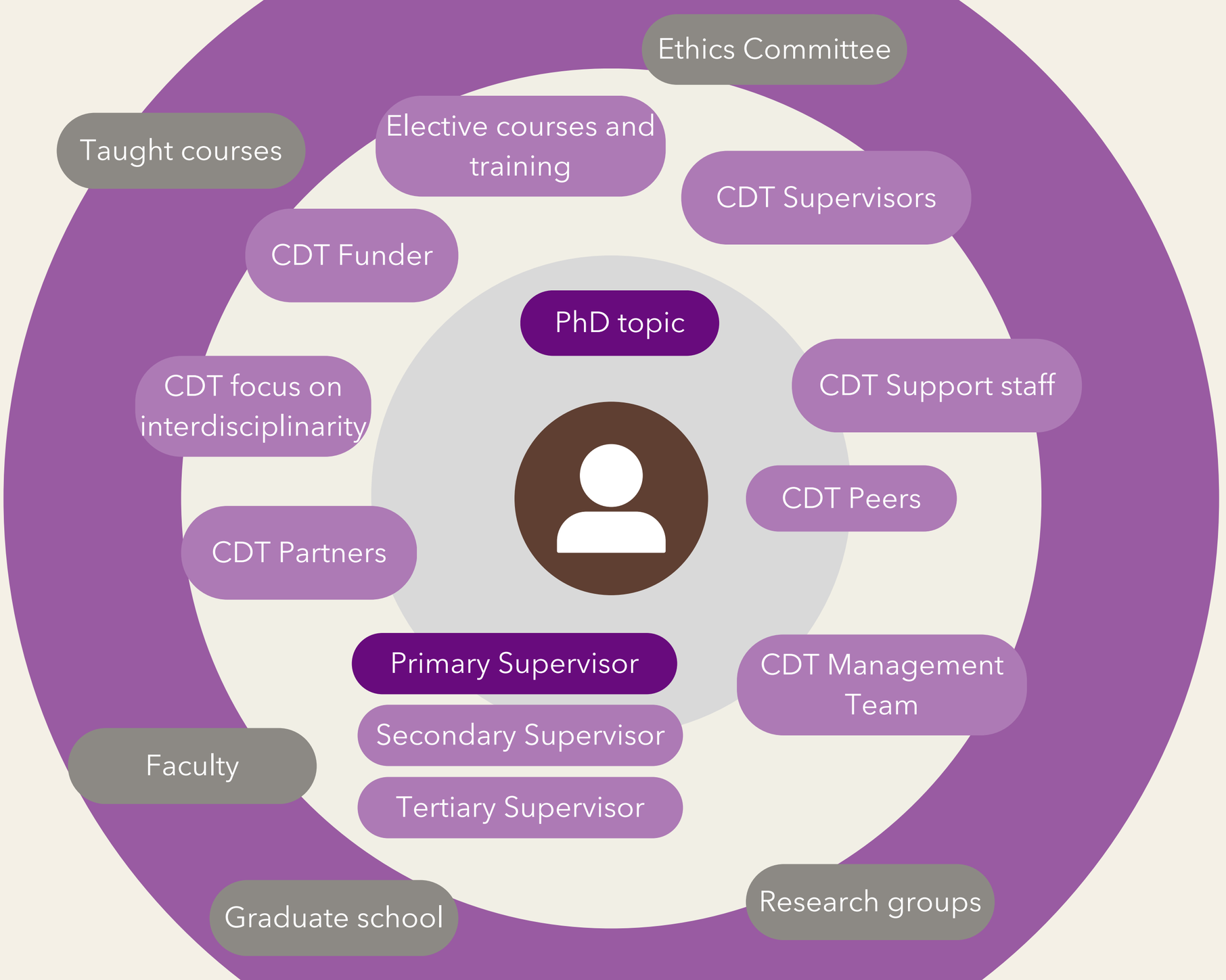
Tom Bidewell

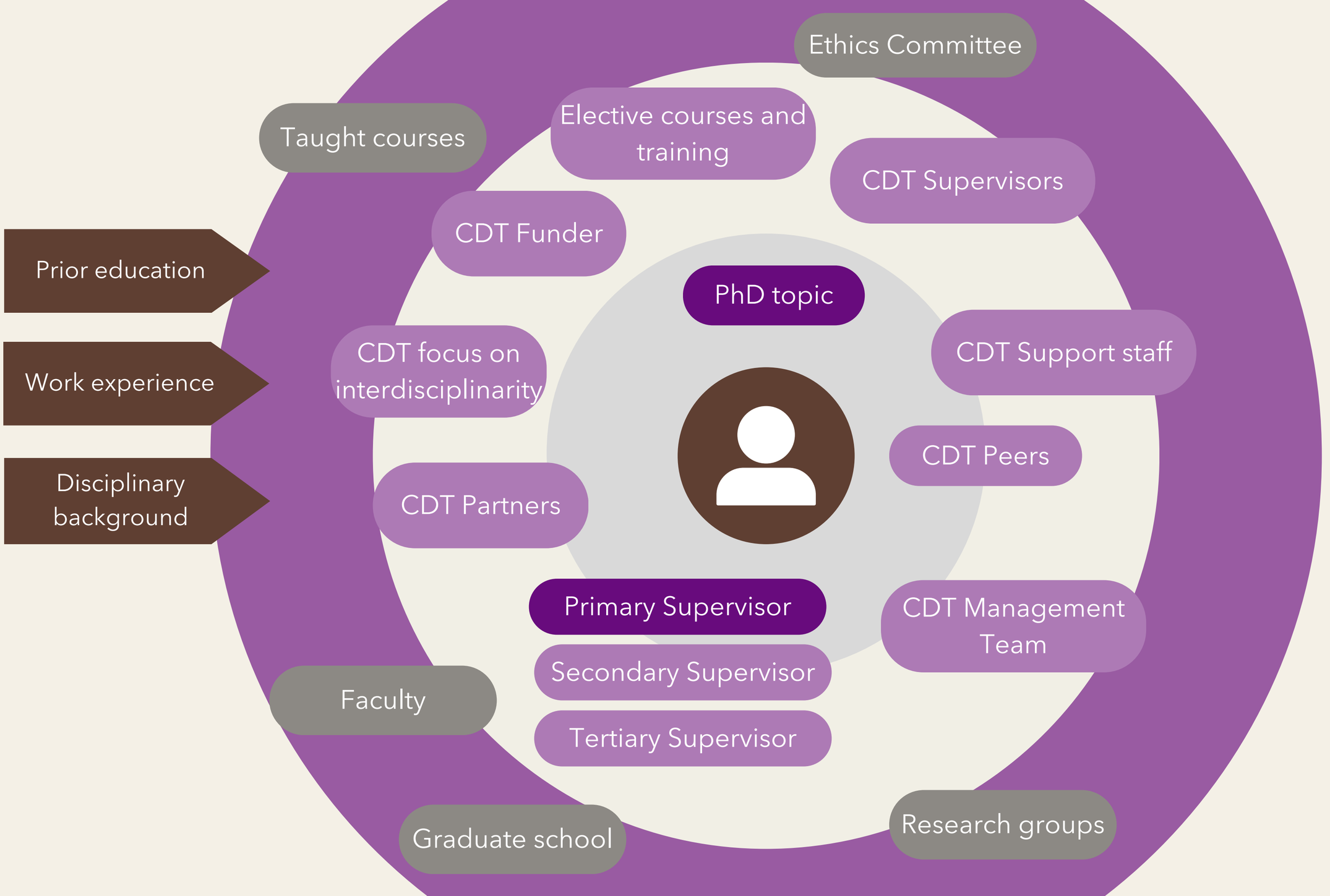
Cohort 1, 2024

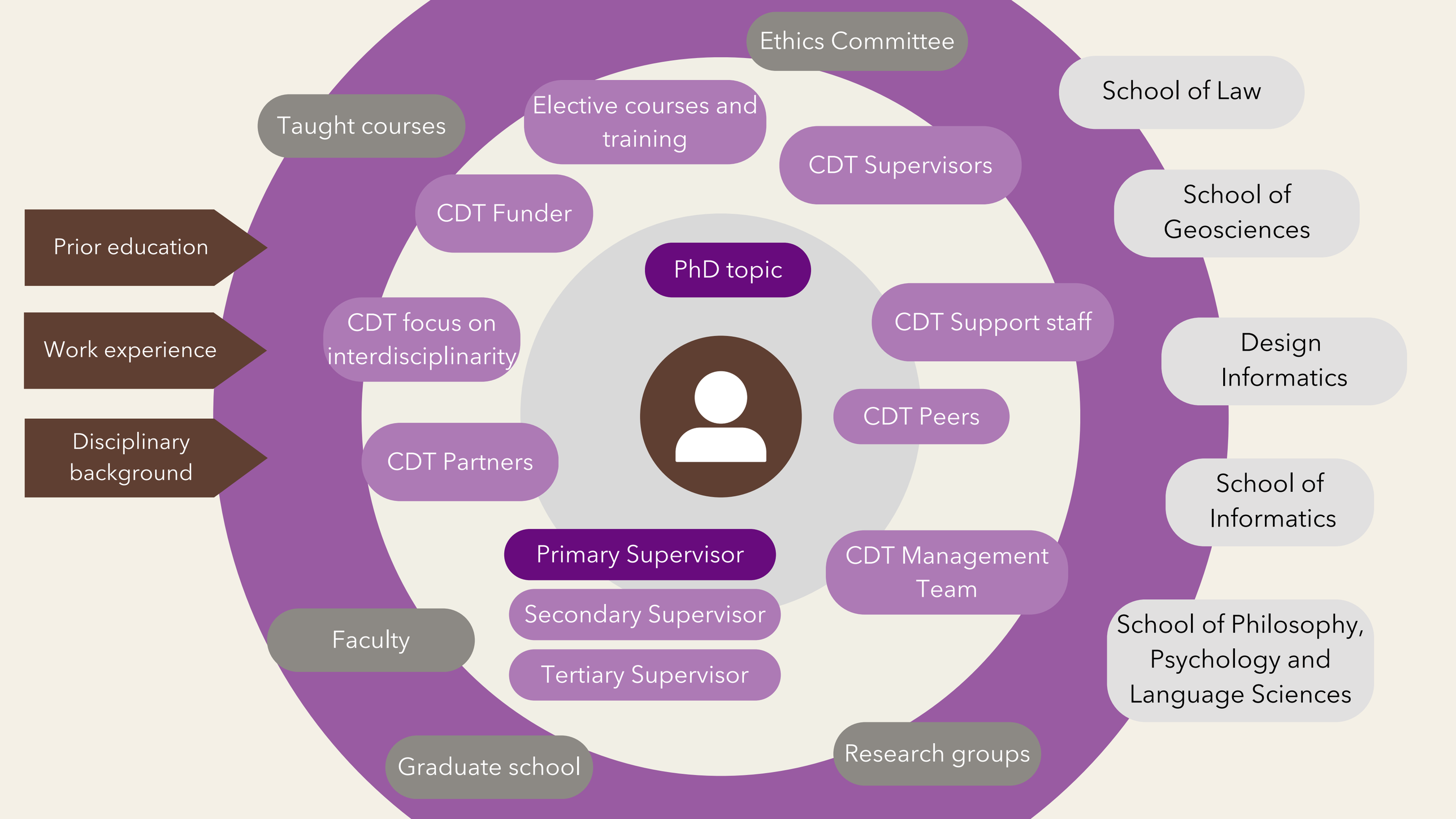

Ethics Committee
Taught courses
PhD topic
Primary Supervisor
Faculty
Graduate school
Research groups


Ethics Committee
Taught courses
PhD topic
CDT Peers
Primary Supervisor
Faculty
Graduate school
Research groups


Ethics Committee
Taught courses
Elective courses and training
CDT Supervisors
CDT Funder
PhD topic
CDT Support staff
CDT Peers
CDT Partners
Primary Supervisor
CDT Management Team
Secondary Supervisor
Faculty
Tertiary Supervisor
Graduate school
Research groups


Ethics Committee
Taught courses
Elective courses and training
CDT Supervisors
CDT Funder
PhD topic
CDT Support staff
CDT focus on interdisciplinarity
CDT Peers
CDT Partners
Primary Supervisor
CDT Management Team
Secondary Supervisor
Faculty
Tertiary Supervisor
Research groups
Graduate school


Ethics Committee
Elective courses and training
Taught courses
CDT Supervisors
CDT Funder
Prior education
PhD topic
CDT Support staff
CDT focus on interdisciplinarity
Work experience
CDT Peers
Disciplinary background
CDT Partners
Primary Supervisor
CDT Management Team
Secondary Supervisor
Faculty
Tertiary Supervisor
Research groups
Graduate school


Ethics Committee
School of Law
Taught courses
Elective courses and training
CDT Supervisors
School of Geosciences
CDT Funder
Prior education
PhD topic
CDT Support staff
Work experience
CDT focus on interdisciplinarity
Design Informatics
CDT Peers
Disciplinary background
CDT Partners
School of Informatics
Primary Supervisor
CDT Management Team
Secondary Supervisor
Faculty
School of Philosophy, Psychology and Language Sciences
Tertiary Supervisor
Research groups
Graduate school

Key *Questions*

In this particular educational and institutional context, and **unique PhD programme**, what **conceptualisations of 'responsible NLP'** dominate?

What actors, institutions and experiences **inform CDT students'** perspectives and decisionmaking on 'responsible' NLP?

What **barriers** do CDT students experience **in implementing** notions of 'responsibility' in practice?

Benedetta Catanzariti
School of Social and Political Science

Fabio Tollon
Edinburgh Futures Institute

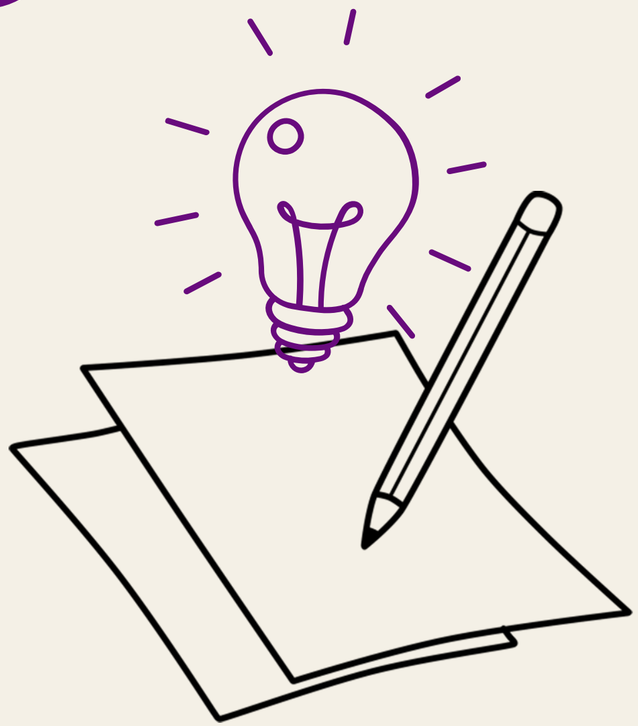
Hannah Rohde
School of Philosophy, Psychology and Language Sciences

James Garforth
School of Informatics

Research *Objectives*

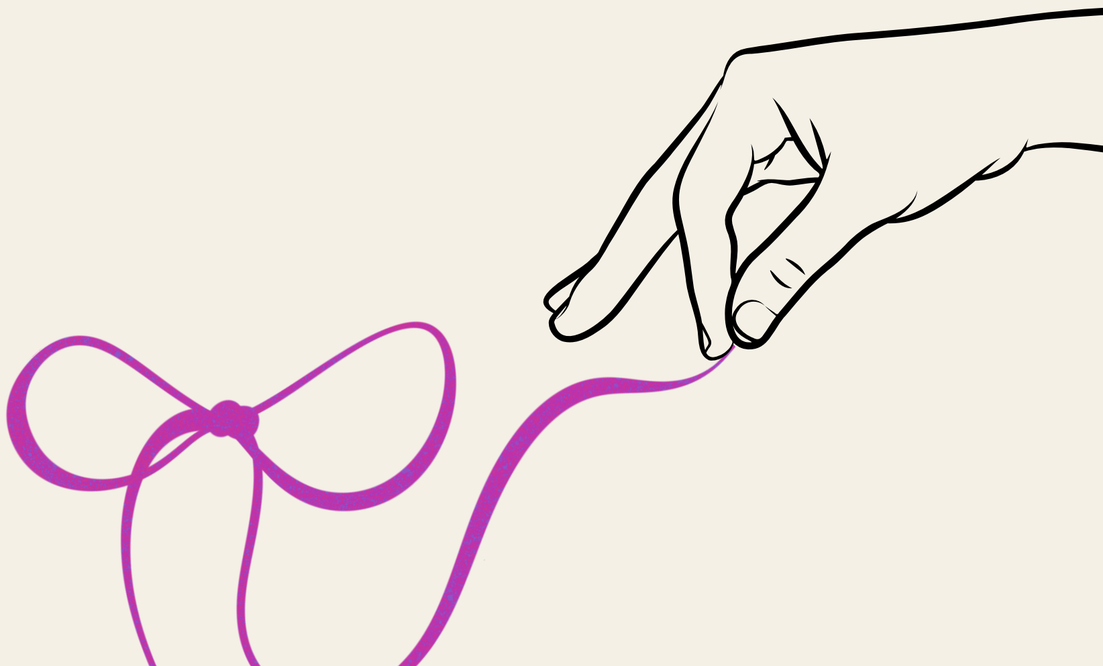
1

2



Study Design

- How to **engage students** meaningfully and **critically**?
- Move from **abstract discussions** into **notions** tested in **practice?**



Study Motifs

- **Definition** of Responsible AI/NLP
- What **influences** their perceptions of Responsible AI/NLP
- What are their **experiences** in relation to the term

Study Structure

Contextualise

Discuss & Test

Reflect

Responsible AI?

What is the **first thing** that comes to mind ?

AI-Incidents

Words

Harms

Responsible AI?

What is the first thing that comes to mind ?

Frameworks

Actors

Use-Cases

Outcomes

Study Design: Methods

Survey Design

- **What voices** are we capturing?
- **Identify alignment** against the **length of institutional affiliation**
- **Conceptualising** the term using **words** and **examples**
- **Relevance** of Responsible AI/NLP in their work
- Subjective Evaluation of **Accountability of Actors** in the AI Supply Chain
- **Familiarity** with Ethical and Legal **Frameworks** on Responsible AI
- **Evaluate opinions** around LLM-related activities and situations
- **Sources of Influence** of various groups and institutions on perception

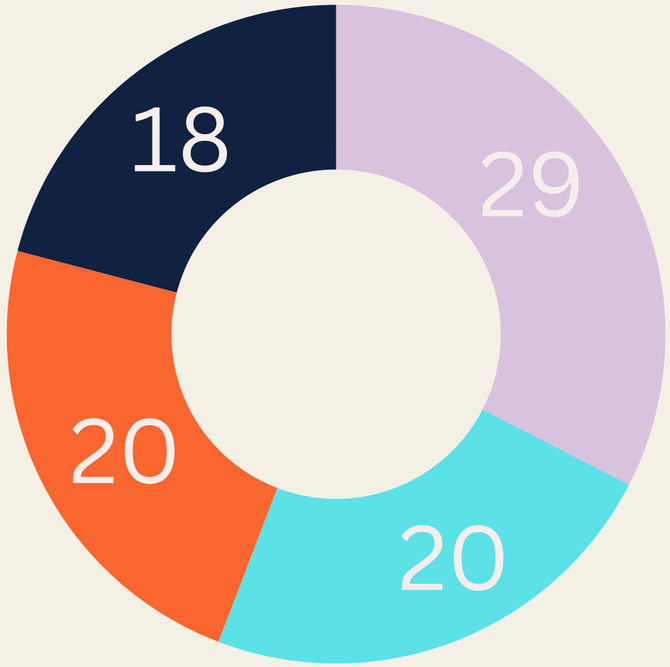
Study Design: Survey

Staff

Postgrad (resear...

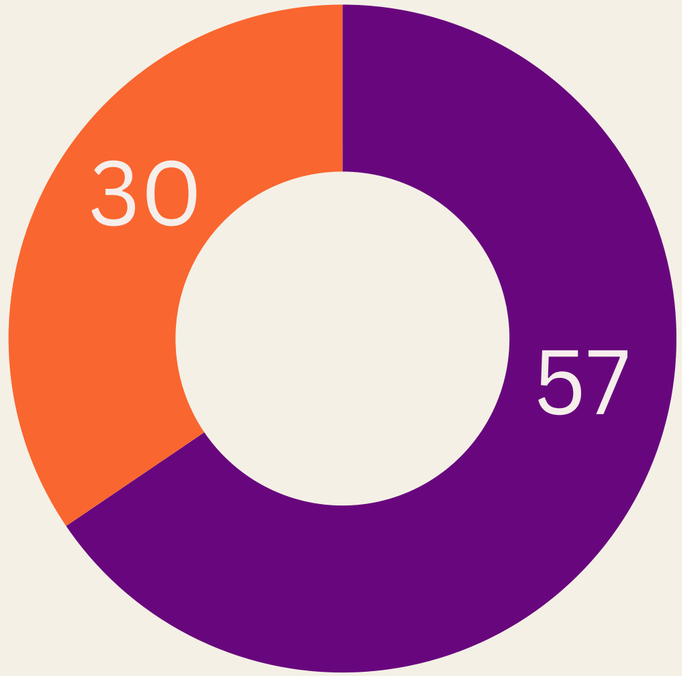
Postgrad (taught)

Undergrad



School of Inform...

Other school

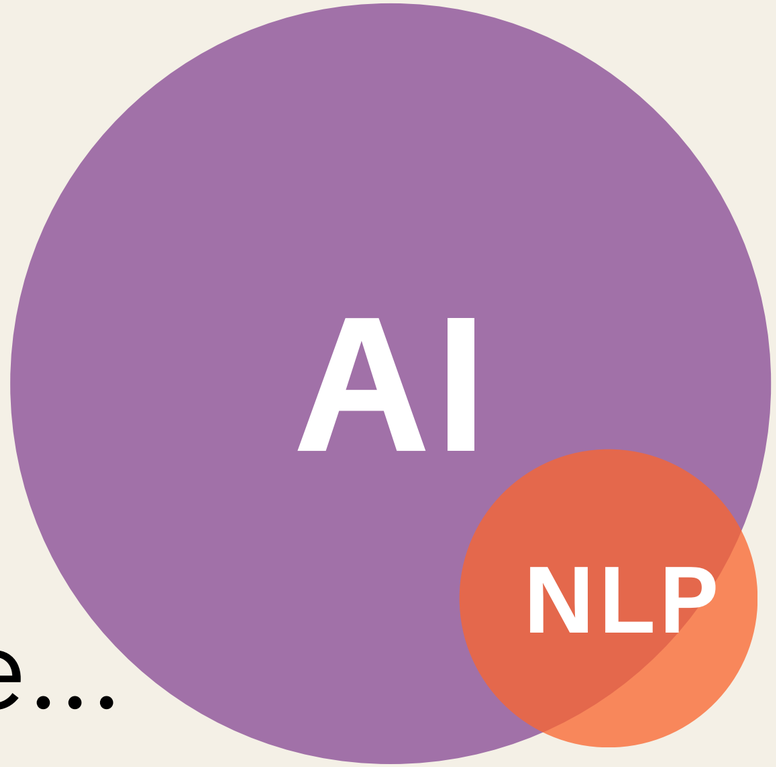


Respondant Demographics

- **86** Respondents
- **24** CDT Staff, Students

Segmenting Perspectives

Responsible NLP CDT vs **Other Groups @University**



Responsible...

Study Design: Survey

Responsible AI?

Any **three words** that come to mind ?

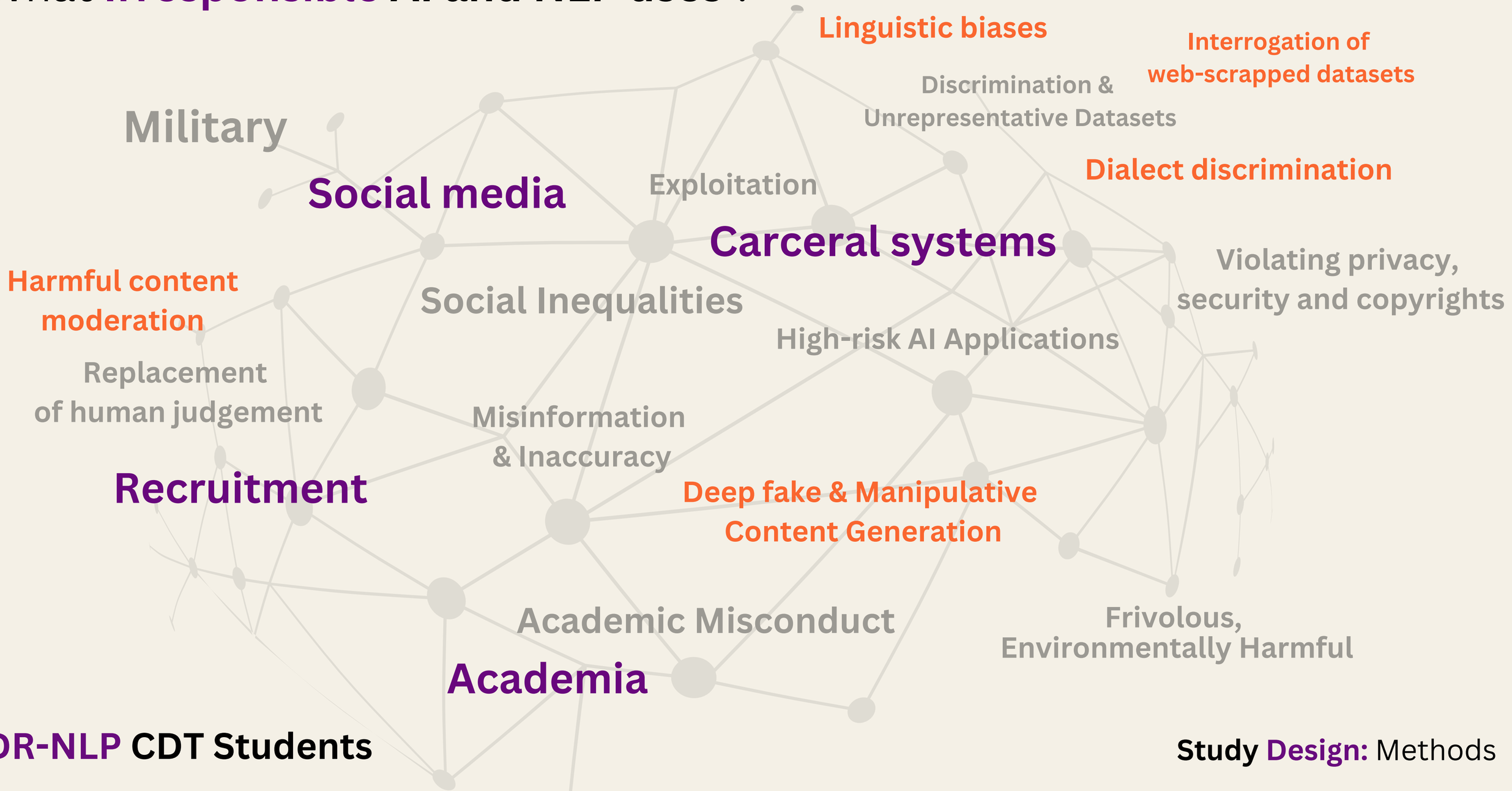
Any **three words** that come to mind ?

What **Irresponsible** AI and NLP uses ?

University Network

Study Design: Methods

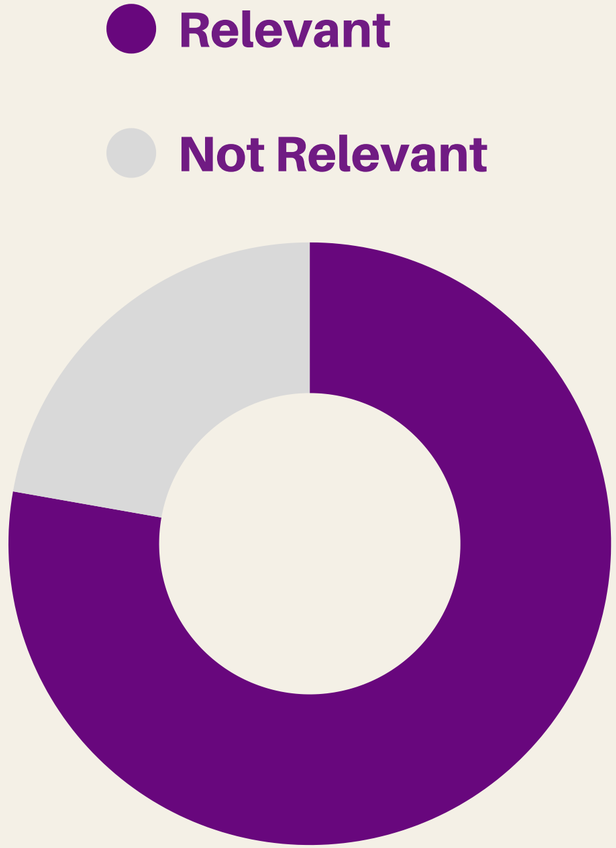
What **Irresponsible** AI and NLP uses ?



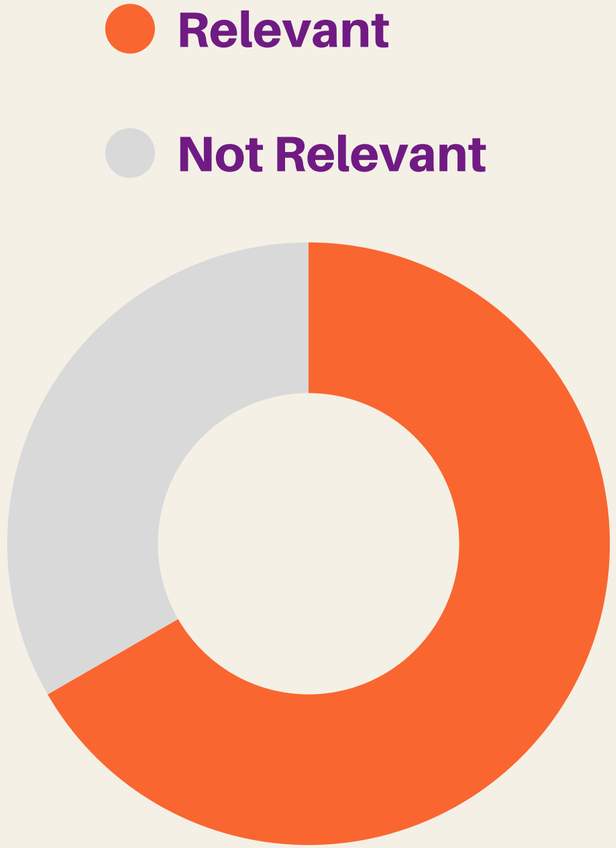
DR-NLP CDT Students

Study Design: Methods

Responsible AI is...



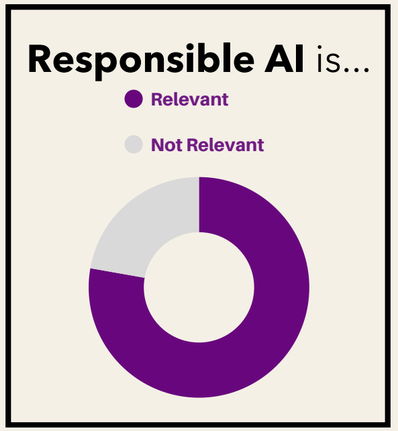
Responsible NLP is...

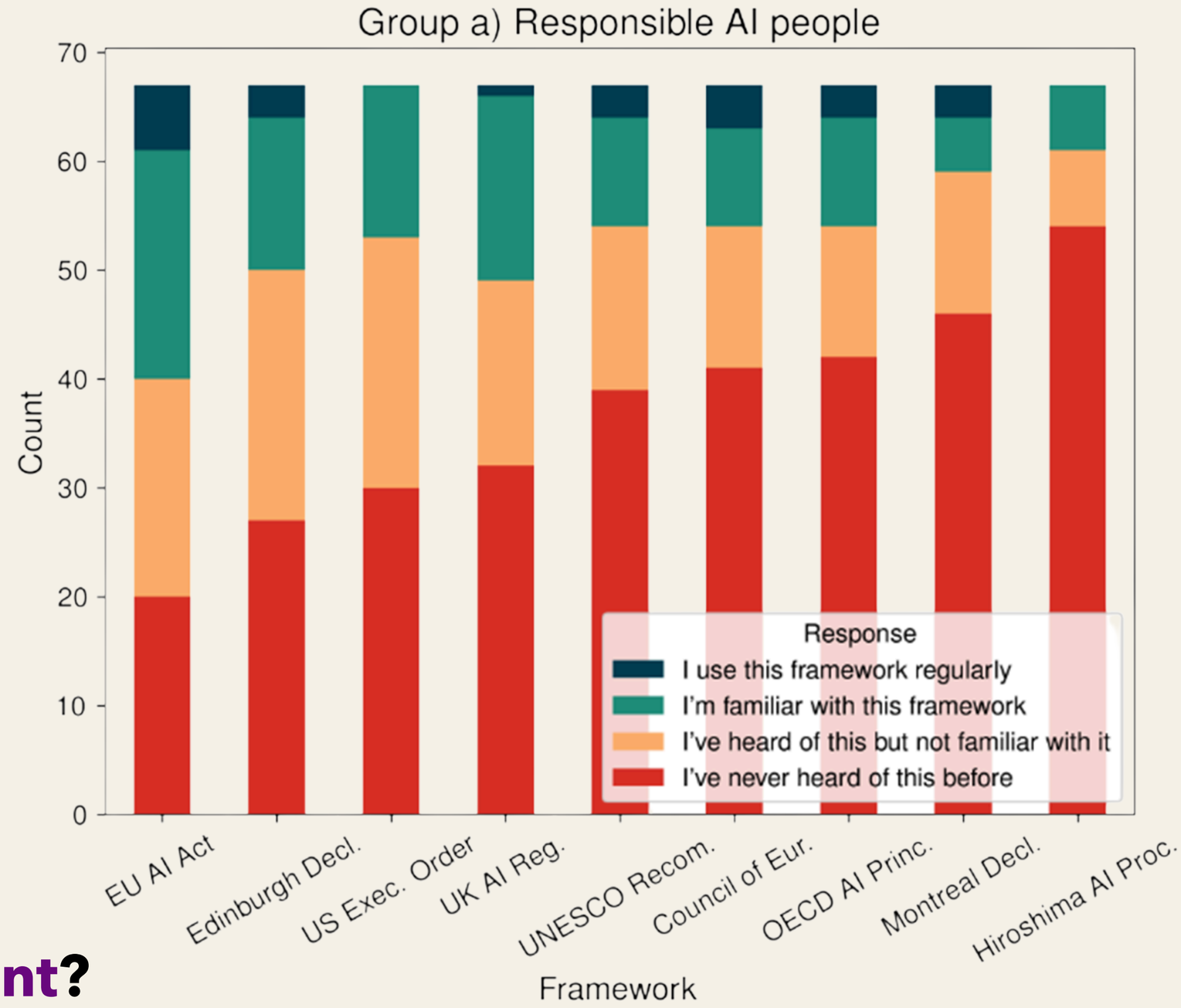


...to day-to-day research and work

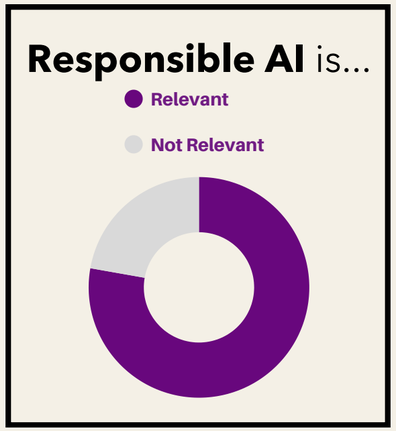
How relevant?

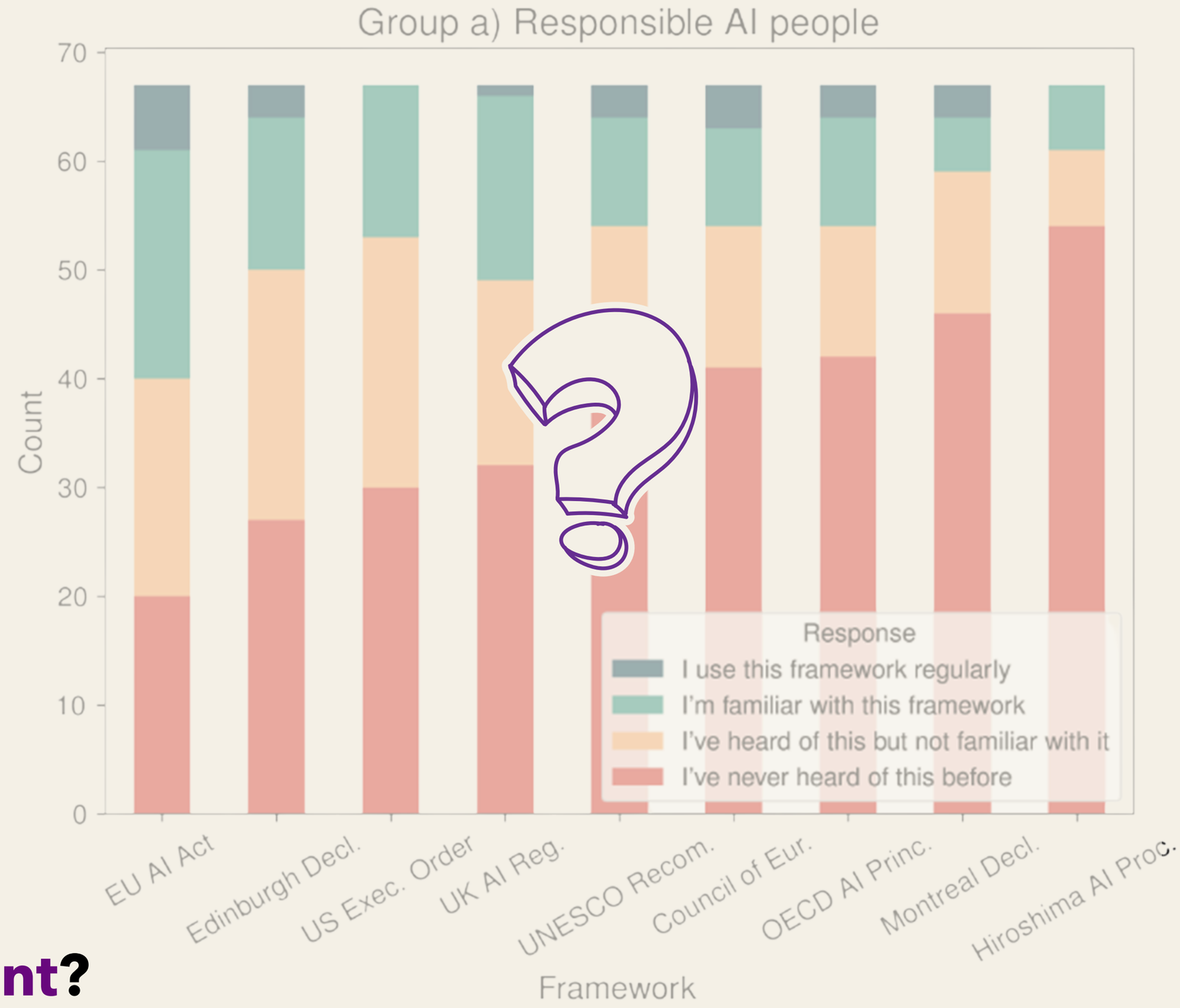
How relevant?



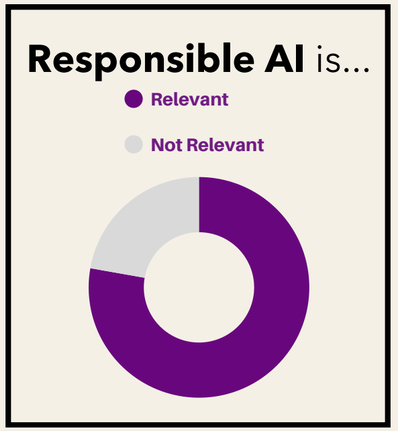


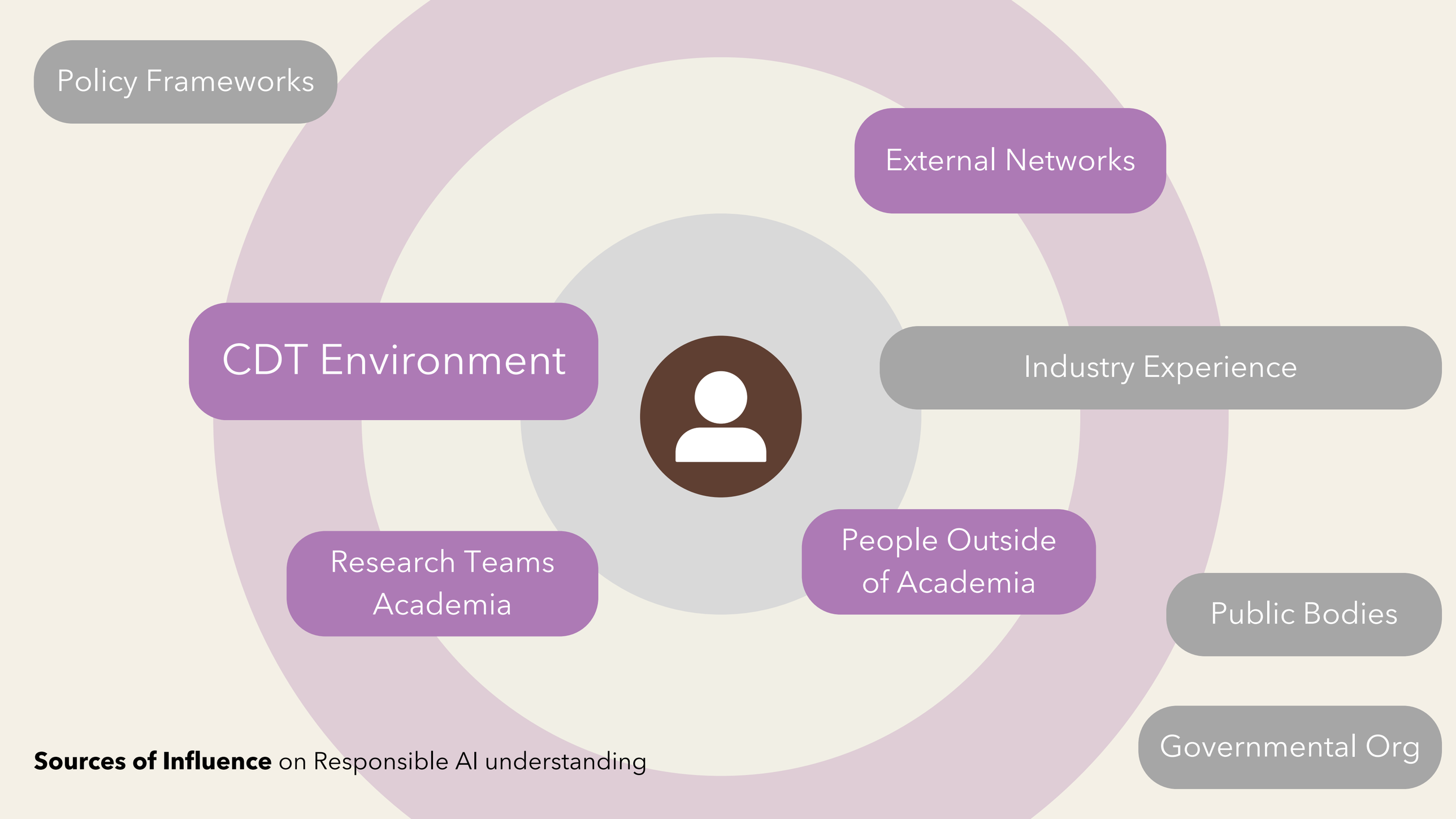
How relevant?





How relevant?





Sources of Influence on Responsible AI understanding

Workshop Design

- **ICE BREAKER**
- **OPEN DISCUSSION with SILENT REFLECTION**
- **PROJECT DEVELOPMENT PIPELINE**

Workshop Design

- **ICE BREAKER** (5 min each)

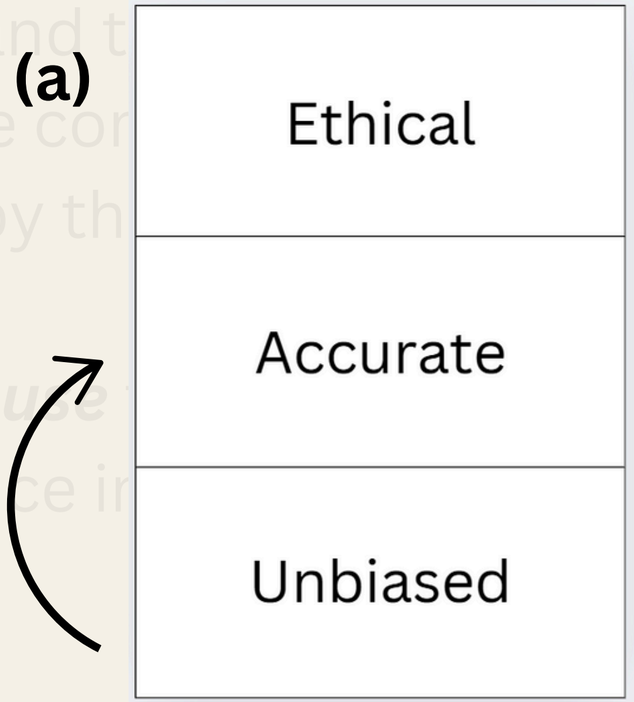
  Understand the ***importance assigned* to words**, by CDT students, which are commonly associated words with Responsible AI that were elected by the wider institutional community

  Not only ***use*** the terms, but ***critically assess their form***, meaning and importance in relation to one another.

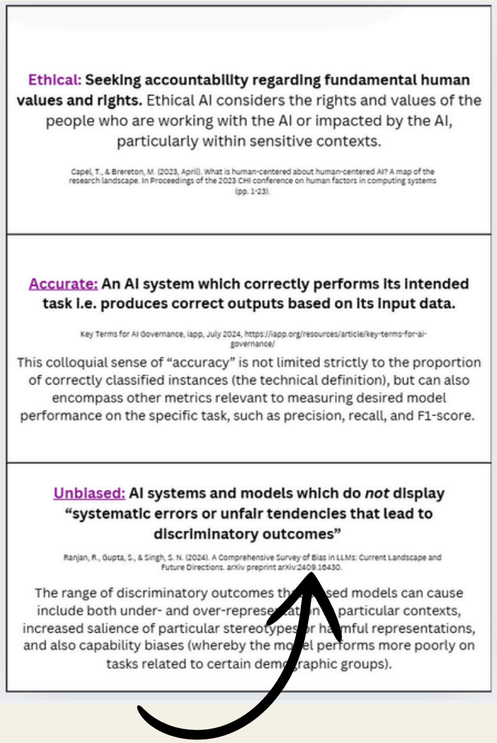
Workshop Design

- **ICE BREAKER**

(a)



(b)



Study Design: Survey

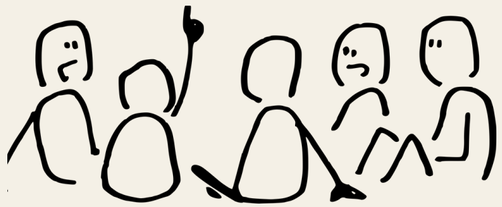
Ranking Activity

Workshop Design

- **OPEN DISCUSSION** (6-8 minutes)

  Enable students to **form and reflect on their own understanding** of the question and formulate answers *before* building on each other's comments in an open discussion.

  The open discussion serves to **capture collective responses** that reflect *consensus* or *disagreement* of participants.



Workshop Design

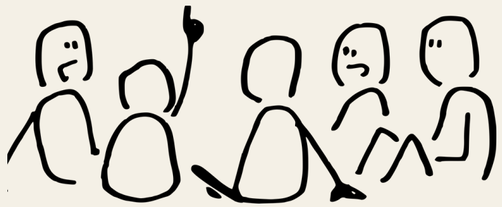
- **OPEN DISCUSSION with SILENT REFLECTION (+ 2 min)**

  Presented with the research questions and relevant prompts

  e.g. How would you define responsible AI/NLP?

  Has your definition changed over time due to a particular experience or situation? How did it change?

  Are there aspects of Responsible AI that you feel are especially important in NLP? If so, why?



Study Design: Workshop

Workshop Design

- **PROJECT DEVELOPMENT PIPELINE** (45 min)

  Move **from abstracted notions** of ethics and principles **to “practice”** by undertaking a hypothetical ethical AI project based on a real-life scenario.

  Create an **interactive reflective activity** to initiate *constructive discussions* on Responsible AI approaches in research and development.

Study Design: Workshop

PROJECT DEVELOPMENT PIPELINE

You have 12 months

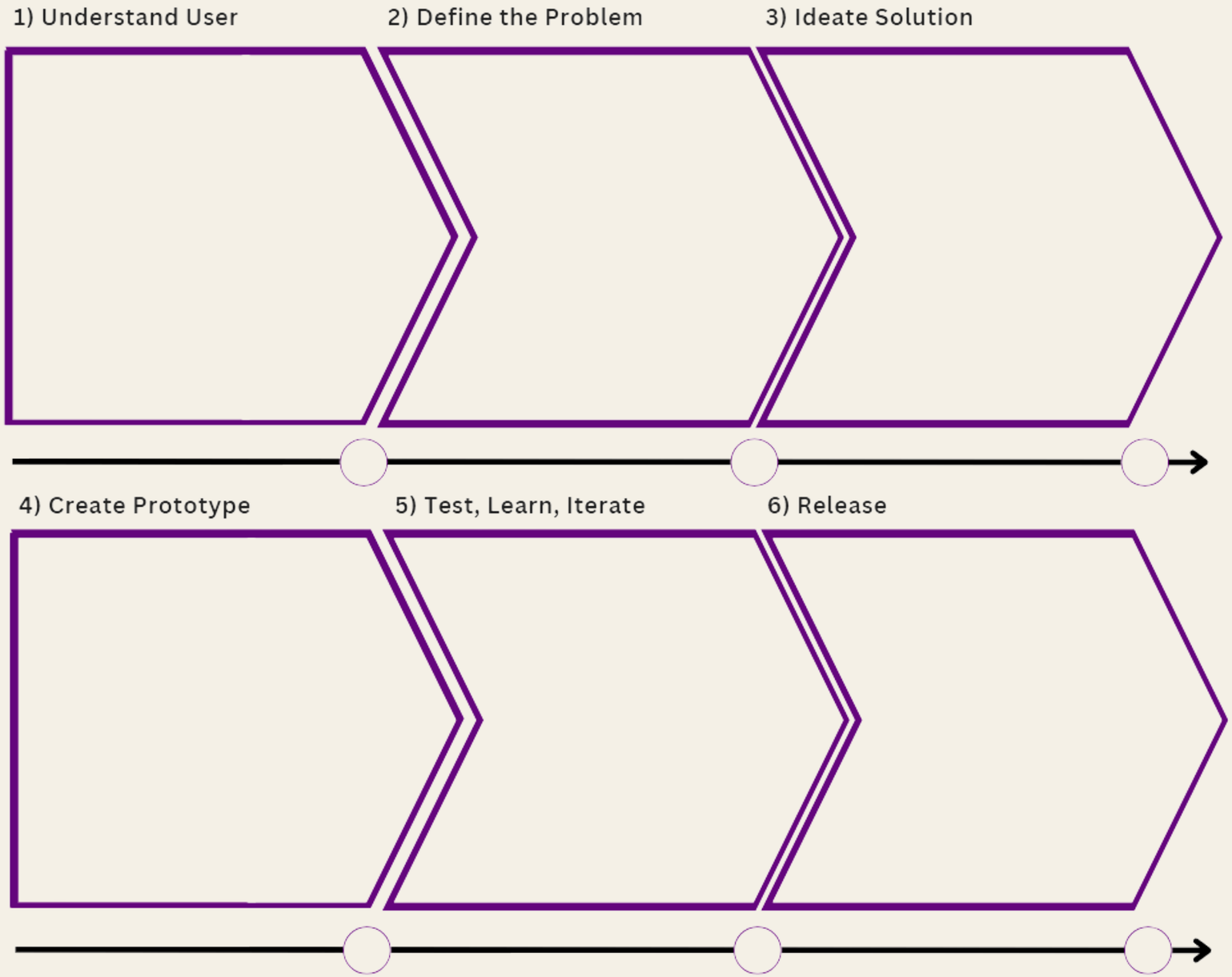
£10 000

Assistive Tech for Deaf Students

Quiz-Question Generation tool for video-based learning

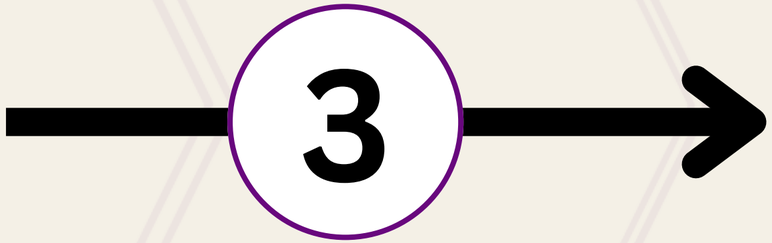
Study Design: Workshop

PROJECT DEVELOPMENT PIPELINE Template



PROJECT DEVELOPMENT PIPELINE

Estimated Time (Months)



Practical Steps
Milestones

Risks

Responsible AI
Measures

PROJECT DEVELOPMENT PIPELINE Considerations

DELIVERING THE PRODUCT

a) **Objectives**, Unique Selling Point, Purpose?

b) **Methods**

c) **Data Collection**, **Model Training**, Deployment, Post-Market Monitoring?

d) **Resource allocation (time, money)**, when/where does responsible AI kick in?

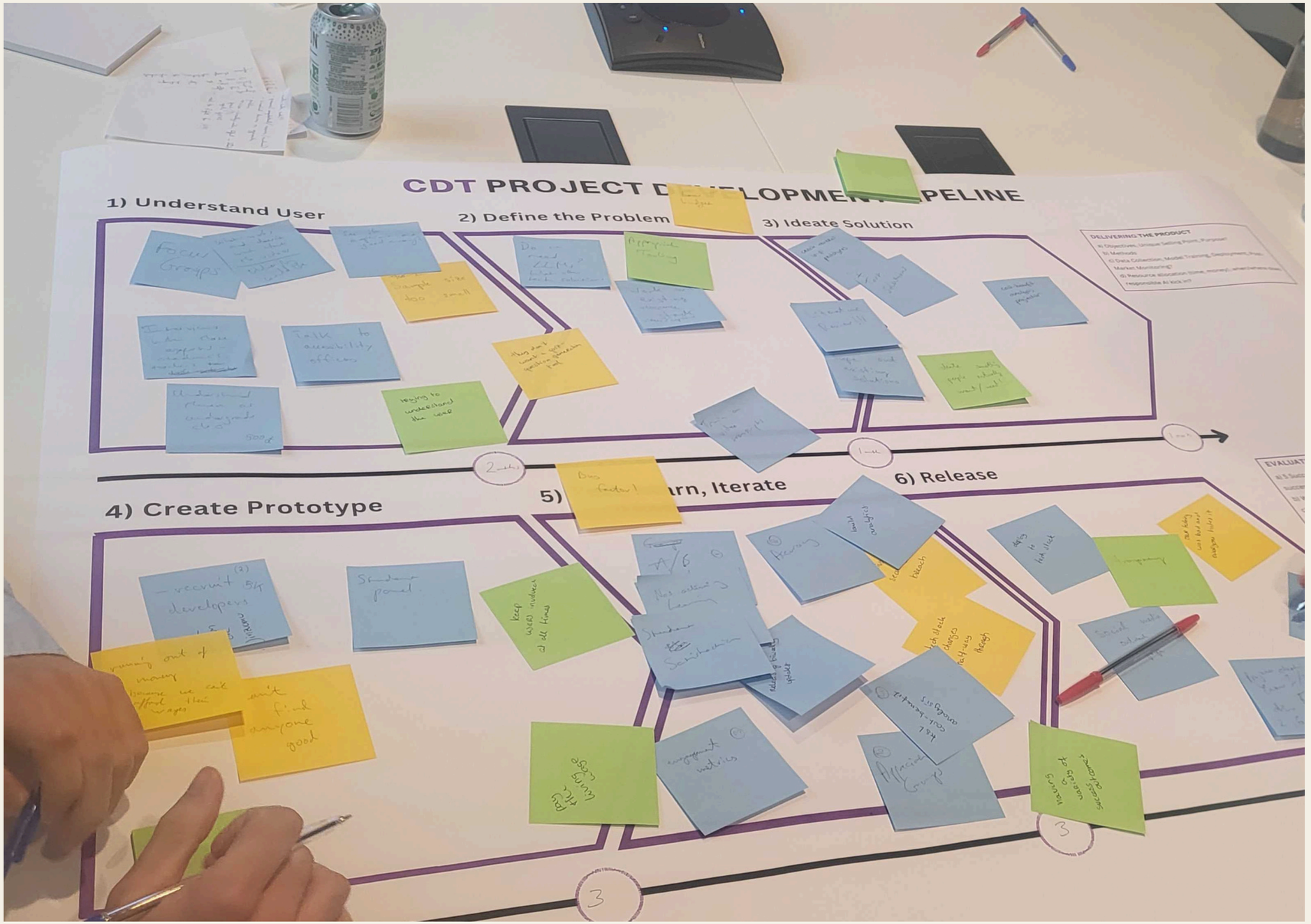
PROJECT DEVELOPMENT PIPELINE Considerations

EVALUATING THE PRODUCT

a) **5 Success Metrics** (how will you know your model is successful post-deployment?), Expected Accuracy?

b) What are your **guardrails**? What are the **limitations**?

c) Within your CDT environment, consider what is the **role of CDT Management** Staff, University, Other Students, Supervisors, Partners


CDT PROJECT D LOPMEN PELINE
1) Understand User
2) Define the Problem
3) Ideate Solution
4) Create Prototype
5) rn, Iterate
6) Release
DELIVERING THE PRODUCT
Focus Groups
Talk to accessibility officer
trying to understand the user
keep users involved all the time
2 wks
1 wk
3

PROJECT DEVELOPMENT PIPELINE

RUG PULL-ELEMENT

3 months

You have ~~12 months~~

~~£10 000~~ **£2000**

Assistive Tech for Deaf Students

Quiz-Question Generation tool for video-based learning

Workshop Design

Background:

- Ethical work, critical thinking and reflexive practice are often **non-existent or siloed** in computer science curricula (Catanzariti, 2024; Moats, 2019; Raji, 2021; Darling-Wolf, 2024)...although some are working to address this (e.g. Reich et al, 2020)

Background:

- Ethical work, critical thinking and reflexive practice are often **non-existent or siloed** in computer science curricula (Catanzariti, 2024; Moats, 2019; Raji, 2021; Darling-Wolf, 2024)...although some are working to address this (e.g. Reich et al, 2020)

- Individuals working in data science and AI often **don't perceive themselves as having agency or responsibility** (Widder & Nafus, Drage, McInerney & Browne, 2024, Sarder, 2022)

Background:

- Ethical work, critical thinking and reflexive practice are often **non-existent or siloed** in computer science curricula (Catanzariti, 2024; Moats, 2019; Raji, 2021; Darling-Wolf, 2024)...although some are working to address this (e.g. Reich et al, 2020)

- Individuals working in data science and AI often **don't perceive themselves as having agency or responsibility** (Widder & Nafus, Drage, McInerney & Browne, 2024, Sarder, 2022)

- There is a **gap** between high-level principles on responsible AI and technological practice (Munn, 2022)...translation from theory to practice takes work (Tanweer, 2022; Stone et al, 2023).

Our findings

Our findings

CDT students feel strongly that ‘responsible AI’ ...

- is bigger than just the models

Our findings

CDT students feel strongly that ‘responsible AI’ ...

- **is bigger than just the models**

> *These systems aren't just responsible for doing a task.* ***They need to be responsible to everyone they affect.***

Our findings

CDT students feel strongly that ‘responsible AI’ ...

- is bigger than just the models
- **includes letting people opt out**

Our findings

CDT students feel strongly that ‘responsible AI’ ...

- is bigger than just the models
- **includes letting people opt out**

> *“how are people able to say, no, either **I'm not okay with the way that this model works** or what it's doing or what it's collecting or **I'm not okay with this application in a public space**.”*

Our findings

CDT students feel strongly that ‘responsible AI’ ...

- is bigger than just the models
- includes letting people opt out
- **is not techno-solutionism**

Our findings

CDT students feel strongly that ‘responsible AI’ ...

- is bigger than just the models
- includes letting people opt out
- **is not techno-solutionism**

> *I feel like there's **a big push for using immature technologies for sensitive sectors** where the capabilities just aren't there.*

Our findings

CDT students feel strongly that ‘responsible AI’ ...

- is bigger than just the models
- includes letting people opt out
- is not techno-solutionism
- **is linked to financial incentives**

Our findings

CDT students feel strongly that ‘responsible AI’ ...

- is bigger than just the models
- includes letting people opt out
- is not techno-solutionism
- **is linked to financial incentives**

> *[the company that I worked at] showed me that this sort of environment can be so stressful that there* ***you will have people make decisions to cut corners.*** *It's just the reality, which* ***comes back to capitalism.***

Our findings

CDT students perceive barriers to ‘responsible AI’ as...

Our findings

CDT students perceive barriers to ‘responsible AI’ as…

- **Resource constraints within academic contexts**

Our findings

CDT students perceive barriers to ‘responsible AI’ as…

- **Resource constraints within academic contexts or partnerships**

> *I think that the research community has to start funding [more diverse subjects]…because* ***it is noticable that more and more funding has gone through the STEM subjects*** *rather than for instance in linguistics.*

Our findings

CDT students perceive barriers to ‘responsible AI’ as…

- Resource constraints within academic contexts
- **Limitations on what is considered valid academic research**

Our findings

CDT students perceive barriers to ‘responsible AI’ as...

- Resource constraints within academic contexts
- **Limitations on what is considered valid academic research**

> *the fact that these academic benchmarks are still such like a driver of research and* ***it's like benchmark hitting is to me very depressing.***

Our findings

CDT students perceive barriers to ‘responsible AI’ as…

- Resource constraints within academic contexts
- **Limitations on what is considered valid academic research**

> *outside of this academic context, sometimes I feel like the very well-intentioned but so scholarly way of talking about these things* ***starts feeling really fake to me.***

Our findings

CDT students perceive barriers to ‘responsible AI’ as…

- Resource constraints within academic contexts
- **Limitations on what is considered valid academic research**

> *“There's a whole colonial understanding of, like, **what is knowledge? What is research? What is science?** That I don't think aligns at all with when we actually want to do things that are responsible for different communities.”*

Our findings

CDT students perceive barriers to ‘responsible AI’ as…

- Resource constraints within academic contexts
- Limitations on what is considered valid academic research
- **Power dynamics which prevent people saying ‘no’ to irresponsible projects**

Our findings

CDT students perceive barriers to ‘responsible AI’ as…

- Resource constraints within academic contexts
- Limitations on what is considered valid academic research
- **Power dynamics which prevent people saying ‘no’ to irresponsible projects**

> *…that’s not really how the real world works.*

Our findings

"...especially as PhD students, when we are pitching our topics and ideas, should I go for the way that ***maximizes my personal career, that gets more papers accepted?*** *Versus actually* ***focusing on the responsible parts...***"

Trade-offs

Our findings

"...especially as PhD students, when we are pitching our topics and ideas, should I go for the way that ***maximizes my personal career, that gets more papers accepted?*** *Versus actually* ***focusing on the responsible parts...***"

"...even if we want to produce responsible AI or responsible AI research, ***we have to have publications, we have to meet certain requirements...***"

Trade-offs

Our findings

CDT students see interdisciplinarity as key to advancing responsible AI

Our findings

CDT students see interdisciplinarity as key to advancing responsible AI

- Talking to people outside of STEM
- Conversations with people with different perspectives in their masters programme
- Sustained relationships with non-academic communities
- Discussions with those impacted by AI tools and systems in practice
- Reading papers from a different discipline as part of a CDT course
- Opportunities for conversations, training and collaborations afforded by the CDT programme.

Future work

- What types of support for ‘interdisciplinarity’ are most helpful for overcoming barriers to responsible AI / NLP?
- Repeating the study with the next cohort(s) of Designing Responsible NLP CDT students
- Expanding to other computing CDTs
- Investigate with relevant groups in other institutions
  - Let us know if you’re interested in using our materials!

Thank you!

Questions?

Access our survey and workshop materials here

“So, like, here we all are in our, like, tiny little niche of trying to make AI responsible. It's going to take, like, thousands of us to maybe, like, join all those niches up together.”