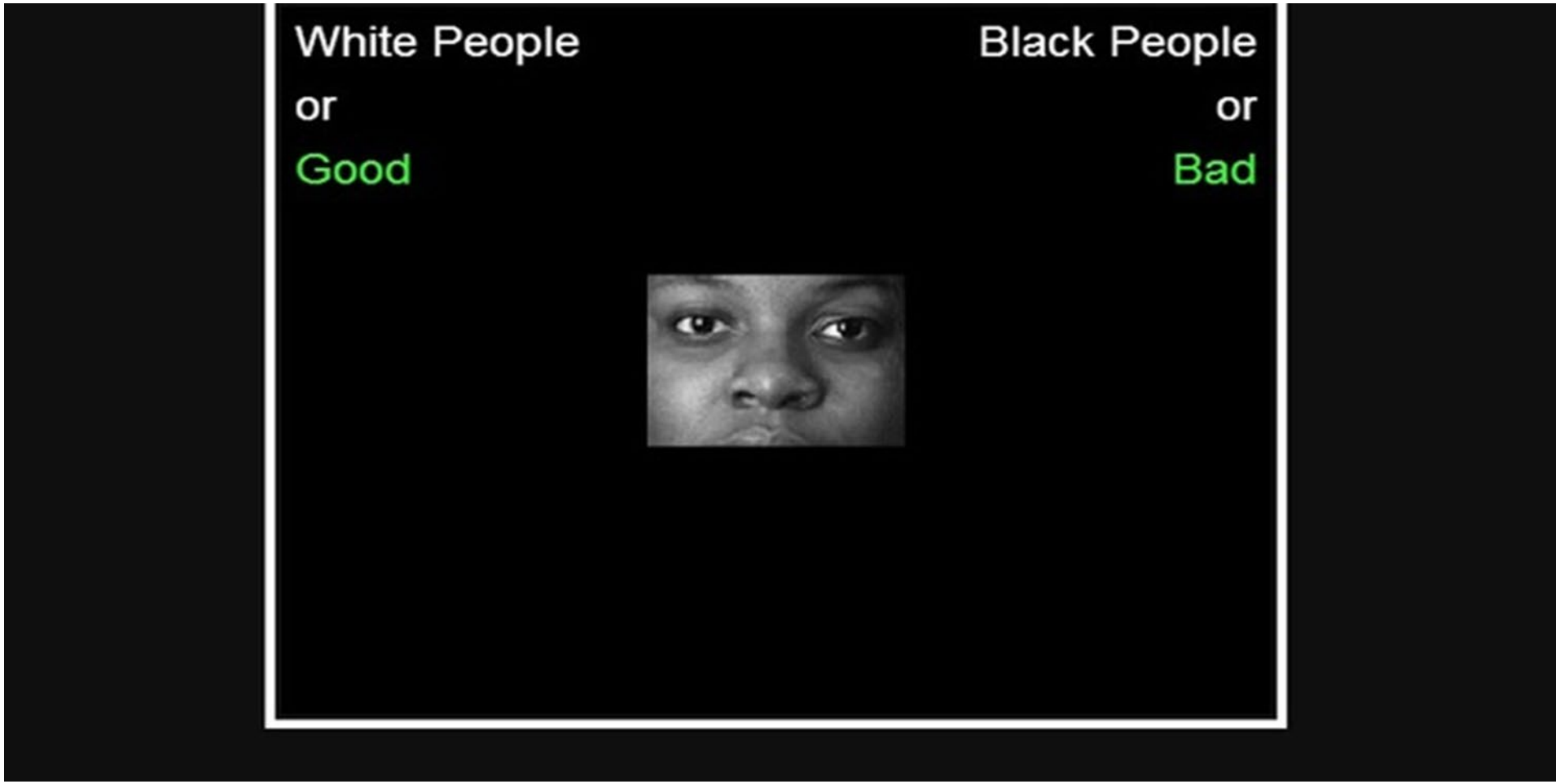
Implicit Bias and Discrimination

Katharina Berndt Rasmussen

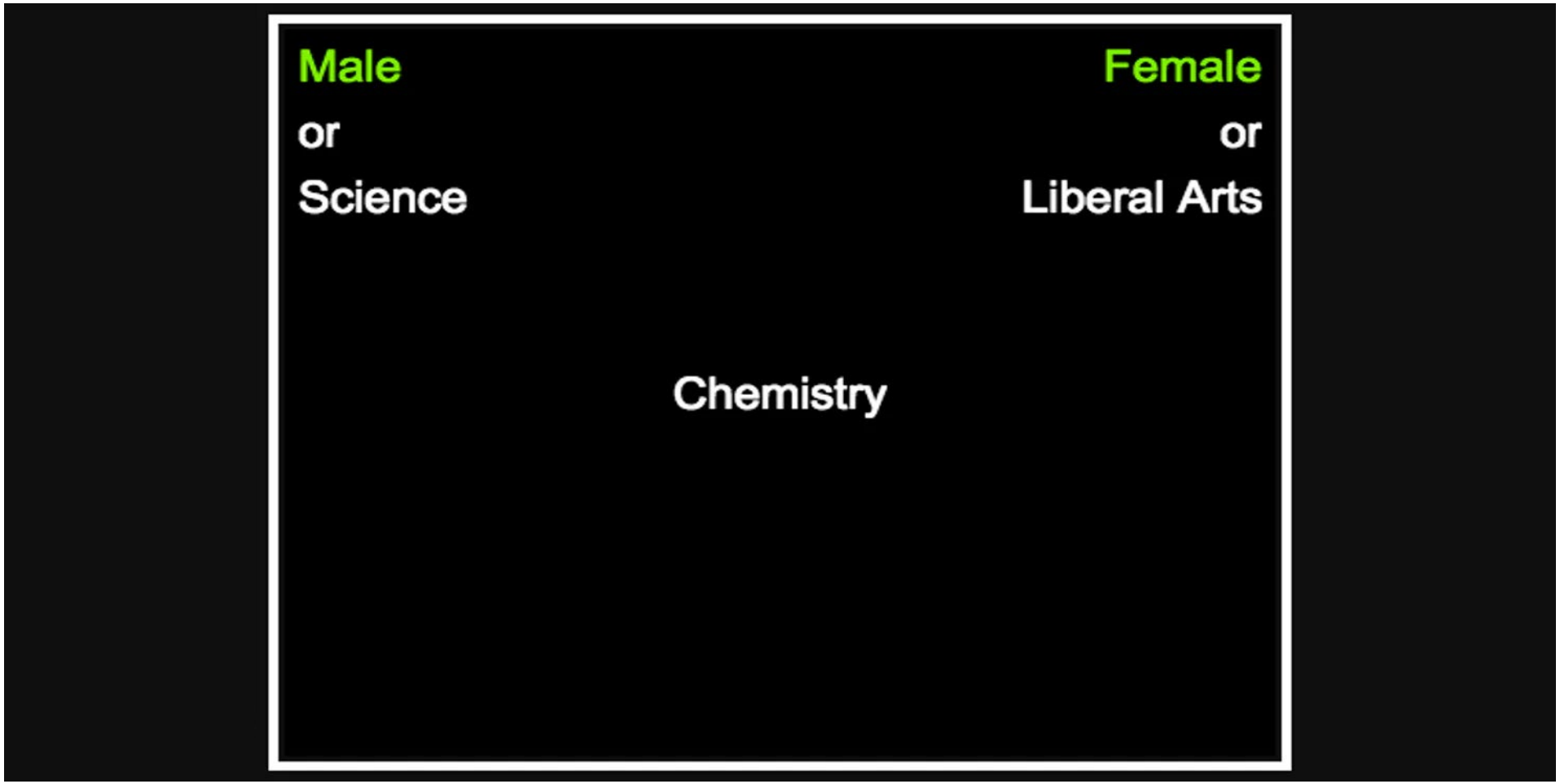
Stockholm University & Institute for Futures Studies

katharina.berndt@iffs.se

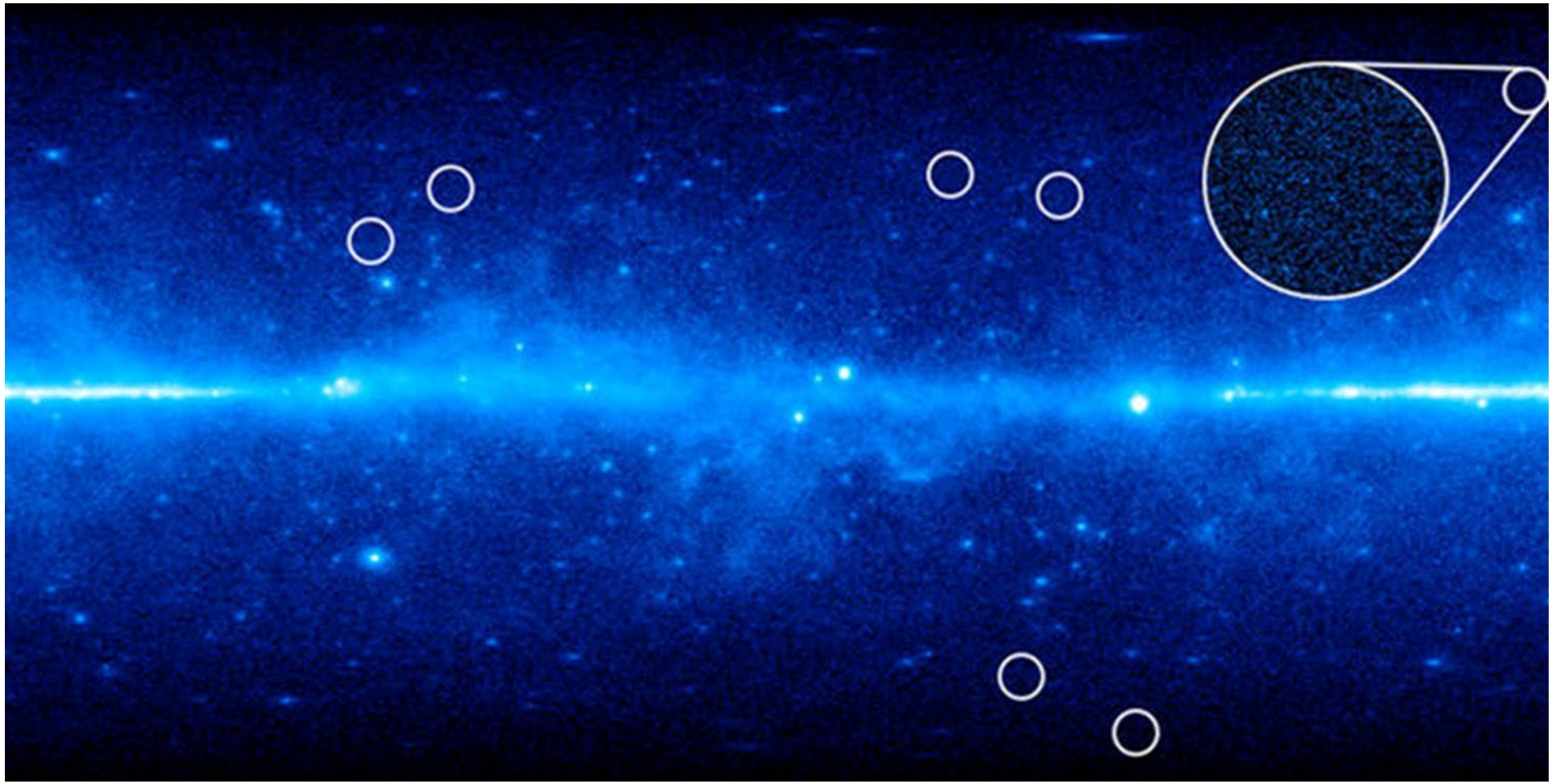
Implicit bias: Implicit Association Test (IAT)



Implicit bias: Implicit Association Test (IAT)



Implicit bias: Human dark matter



Implicit bias

Working definition: implicit biases are whatever mental processes that

i. appear “alien” to their host (unendorsed, not under direct control, unconnected to their behaviour),

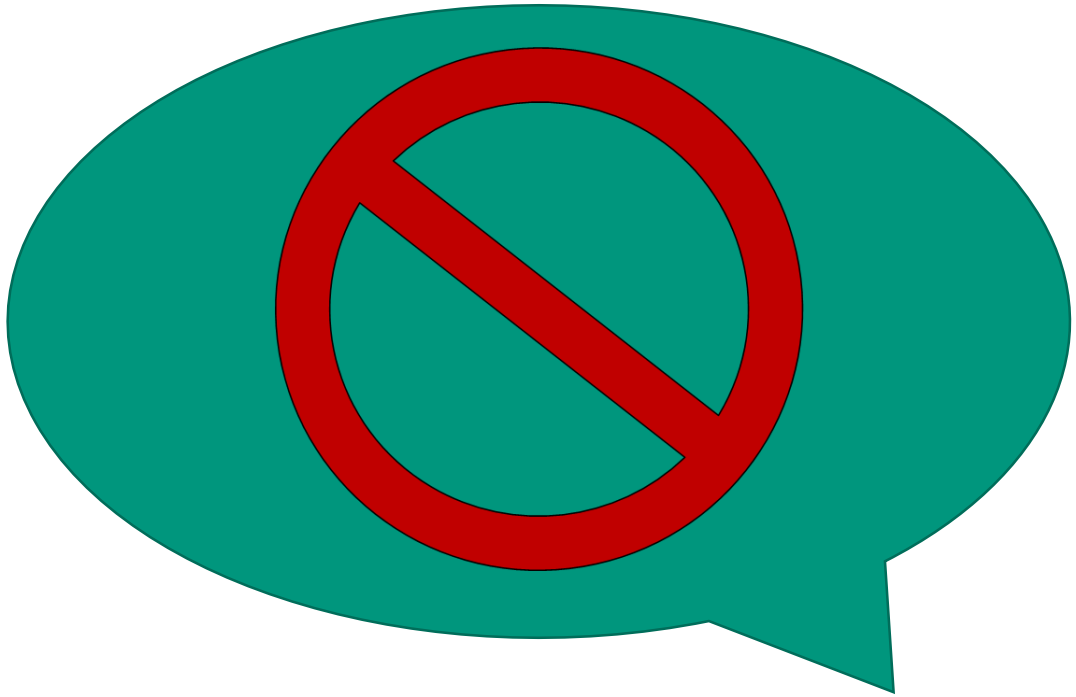
ii. are positively or negatively valenced,

iii. refer to, and are activated by, social categories, and

iv. affect the host’s perceptions, judgments or actions.

- Examples:
  - “Women aren’t good at math.” – “*woman–illogical*”
  - “Black people are dishonest.” – “*Black–dishonest*”


Discrimination!



Discrimination

An agent A discriminates against another individual B when A *treats B worse* – *because* B is a member of a certain group – than A would have treated B, had B not been such a member.

Discrimination: direct vs. indirect

	Disparate treatment	Disparate impact

Discrimination: direct vs. indirect

Intentional		
Non-intentional		

Discrimination: four forms

	Disparate treatment	Disparate impact
Intentional		
Non-intentional		

Discrimination: four forms

	Differential treatment	Disparate impact
Intentional	1) A university in the early 1950's US South accepts a white applicant but turns down an *equally* qualified black applicant, stating: "This is a whites-only university. Blacks are referred to apply to some 'separate-but-equal' university for African Americans."	
Non-intentional		4) An employer turns down a *qualified* black applicant, stating: "We don't hire people who lack high school education", without awareness of the criterion's ability to track politically induced, race-correlated educational deficits.

Discrimination: four forms

	Differential treatment	Disparate impact
Intentional	1) A university in the early 1950's US South accepts a white applicant but turns down an *equally* qualified black applicant, stating: "This is a whites-only university. Blacks are referred to apply to some 'separate-but-equal' university for African Americans."	2) An employer turns down a *qualified* black applicant, stating: "We don't hire people who lack high school education", while intentionally using this criterion because of its ability to track politically induced, race-correlated educational deficits.
Non-intentional		4) An employer turns down a *qualified* black applicant, stating: "We don't hire people who lack high school education", without awareness of the criterion's ability to track politically induced, race-correlated educational deficits.

Discrimination: four forms

	Differential treatment	Disparate impact
Intentional	1) A university in the early 1950's US South accepts a white applicant but turns down an *equally* qualified black applicant, stating: "This is a whites-only university. Blacks are referred to apply to some 'separate-but-equal' university for African Americans."	2) An employer turns down a *qualified* black applicant, stating: "We don't hire people who lack high school education", while intentionally using this criterion because of its ability to track politically induced, race-correlated educational deficits.
Non-intentional	3) A university accepts a white PhD-candidate but turns down an *equally* qualified black PhD-candidate, stating that the latter was *less* qualified, where the unequitable ranking is due to the evaluators' implicit biases.	4) An employer turns down a *qualified* black applicant, stating: "We don't hire people who lack high school education", without awareness of the criterion's ability to track politically induced, race-correlated educational deficits.

Discrimination: conceptual space for implicit bias discrimination

	Differential treatment	Disparate impact
Intentional	1) A university in the early 1950's US South accepts a white applicant but turns down an *equally* qualified black applicant, stating: "This is a whites-only university. Blacks are referred to apply to some 'separate-but-equal' university for African Americans."	2) An employer turns down a *qualified* black applicant, stating: "We don't hire people who lack high school education", while intentionally using this criterion because of its ability to track politically induced, race-correlated educational deficits.
Non-intentional	3) A university accepts a white PhD-candidate but turns down an *equally* qualified black PhD-candidate, stating that the latter ~~was *less*~~ qualified, where the unequitable ranking is due to the evaluators' implicit biases.	4) An employer turns down a *qualified* black applicant, stating: "We don't hire people who lack high school education", without awareness of the criterion's ability to track politically induced, race-correlated educational deficits.

The moral significance – causal connection dilemma

The moral significance – causal connection dilemma

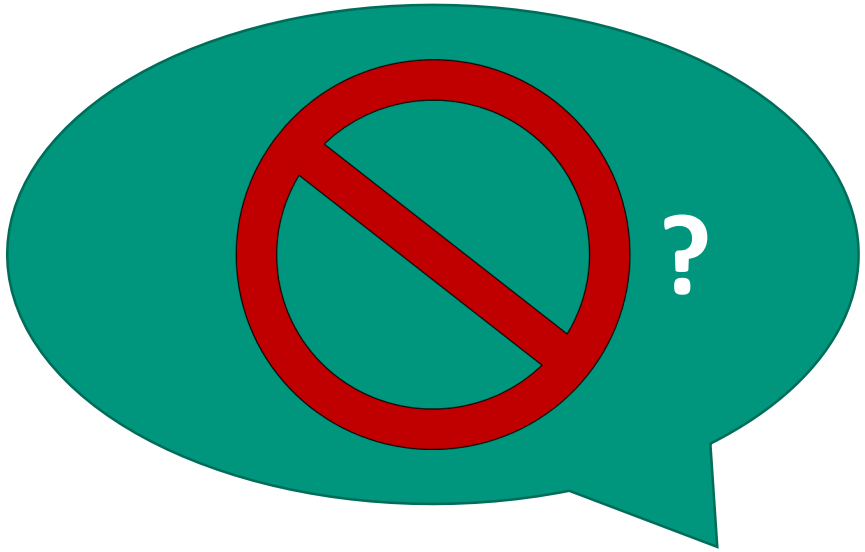
- individual behaviour that is strongly correlated with IAT-scores is (mainly) micro-behaviour that appears morally insignificant
  - examples: smiling frequency, talking time, or eye contact

The moral significance – causal connection dilemma

- individual behaviour that is strongly correlated with IAT-scores is (mainly) micro-behaviour that appears morally insignificant
  - examples: smiling frequency, talking time, or eye contact

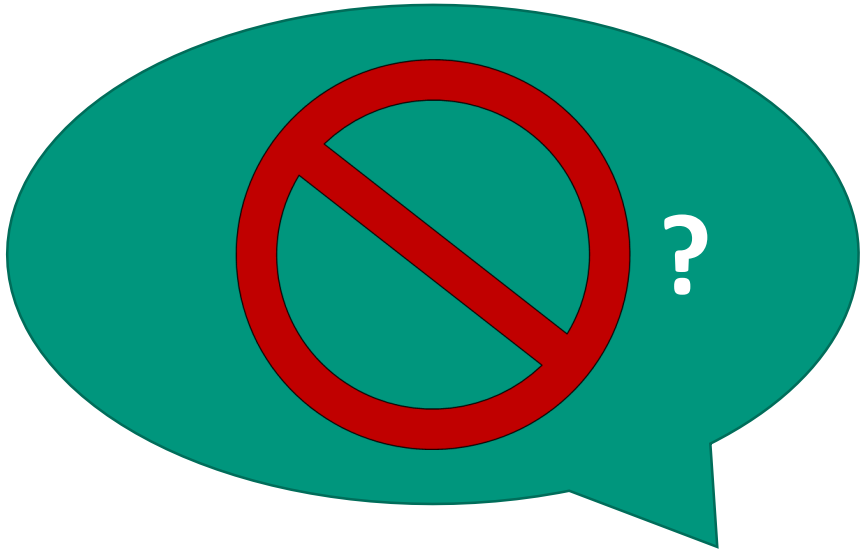
The moral significance – causal connection dilemma

- individual behaviour that is strongly correlated with IAT-scores is (mainly) micro-behaviour that appears morally insignificant
  - examples: smiling frequency, talking time, or eye contact



The moral significance – causal connection dilemma

- individual behaviour that is strongly correlated with IAT-scores is (mainly) micro-behaviour that appears morally insignificant – *not worse treatment in the relevant sense*
  - examples: smiling frequency, talking time, or eye contact



The moral insignificance – causal connection dilemma

- individual behaviour that is strongly correlated with IAT-scores is (mainly) micro-behaviour that appears morally insignificant – *not worse treatment in the relevant sense*
  - examples: smiling frequency, talking time, or eye contact

- morally significant macro-behaviour is only weakly correlated with the individual's IAT-scores
  - example: dismissing an application based on the applicant's race

The moral insignificance – causal connection dilemma

- individual behaviour that is strongly correlated with IAT-scores is (mainly) micro-behaviour that appears morally insignificant – *not worse treatment in the relevant sense*
  - examples: smiling frequency, talking time, or eye contact

- morally significant macro-behaviour is only weakly correlated with the individual's IAT-scores – *not "differential treatment due to implicit bias"*
  - example: dismissing an application based on the applicant's race

The moral insignificance – causal connection dilemma

- individual behaviour that is strongly correlated with IAT-scores is (mainly) micro-behaviour that appears morally insignificant – *not worse treatment in the relevant sense*
  - examples: smiling frequency, talking time, or eye contact

  Yet: micro-behaviours may add up over time and many persons, create e.g. a hostile work environment

- morally significant macro-behaviour is only weakly correlated with the individual's IAT-scores – *not "differential treatment due to implicit bias"*
  - example: dismissing an application based on the applicant's race

The moral insignificance – causal connection dilemma

- individual behaviour that is strongly correlated with IAT-scores is (mainly) micro-behaviour that appears morally insignificant – *not worse treatment in the relevant sense*
  - examples: smiling frequency, talking time, or eye contact

  Yet: micro-behaviours may add up over time and many persons, create e.g. a hostile work environment

- morally significant macro-behaviour is only weakly correlated with the individual's IAT-scores – *not "differential treatment due to implicit bias"*
  - example: dismissing an application based on the applicant's race

  Yet: weak behavioural tendencies may add up over many individuals and rounds of decisions, create patterns of inequality and segregation

Discrimination

An agent A discriminates against another individual B when A *treats B worse* – *because* B is a member of a socially salient group – than A would have treated B, had B not been such a member.

Structural discrimination

A society A structurally discriminates against a group B, whose members share a socially salient property, when under A's societal setup the members of B fare worse than they would have fared, had they not had this socially salient property.

Structural discrimination

A society A structurally discriminates against a group B, whose members share a socially salient property, when under A's societal setup the members of B fare worse than they would have fared, had they not had this socially salient property.

structural implicit bias approach

Structural implicit bias approaches

- *aggregate effects* of individual (strongly IAT-correlated) micro-behaviours or (weakly IAT-correlated) macro-behaviours

- *bias of crowds model*: "implicit bias reflects the accessibility of concepts linked to a social category [the likelihood that a thought, evaluation, stereotype, trait, or other piece of information will be retrieved for use]"; individual implicit bias as "a psychological marker of systemic prejudice in the environment"

- *social norms game-theoretic model*: explicitly egalitarian individual social norm followers occasionally "tremble" and deviate from the norm

Social norms game-theoretic model of implicit bias

- Egalitarian rule: “when hiring, do not discriminate against women or minorities.”
  - People who consider this egalitarian rule a legitimate social norm are conditional rule followers. They prefer to follow the rule if and only if:
    i. *Empirical expectations*: they believe that enough others follow the rule, and
    ii. *Normative expectations*: they believe that enough others expect them to follow the rule.

  - Dynamic game: at the first stage, the players decide what to do based on their beliefs. At the second stage, they update their beliefs (i) and (ii), based on their observations of what the other players did at the first stage. Then they use their new beliefs to decide what to do.
  - Players *sometimes* “tremble” and choose the wrong strategy by mistake: dismiss an application by a qualified female or Black candidate.
  - If the proportion of those who *are observed to* tremble is small enough, the players still believe (i) that enough others follow the rule, and (ii) that enough others expect them to follow the rule, and hence prefer to follow the egalitarian rule next round.

Thank you!

Katharina Berndt Rasmussen

Stockholm University & Institute for Futures Studies

katharina.berndt@iffs.se