Ethics in NLP

Types of risk

Bias amplification: systems exacerbate real-world bias rather than correct for it

Exclusion: underprivileged users are left behind by systems

Dangers of automation: automating things in ways we don’t understand is dangerous

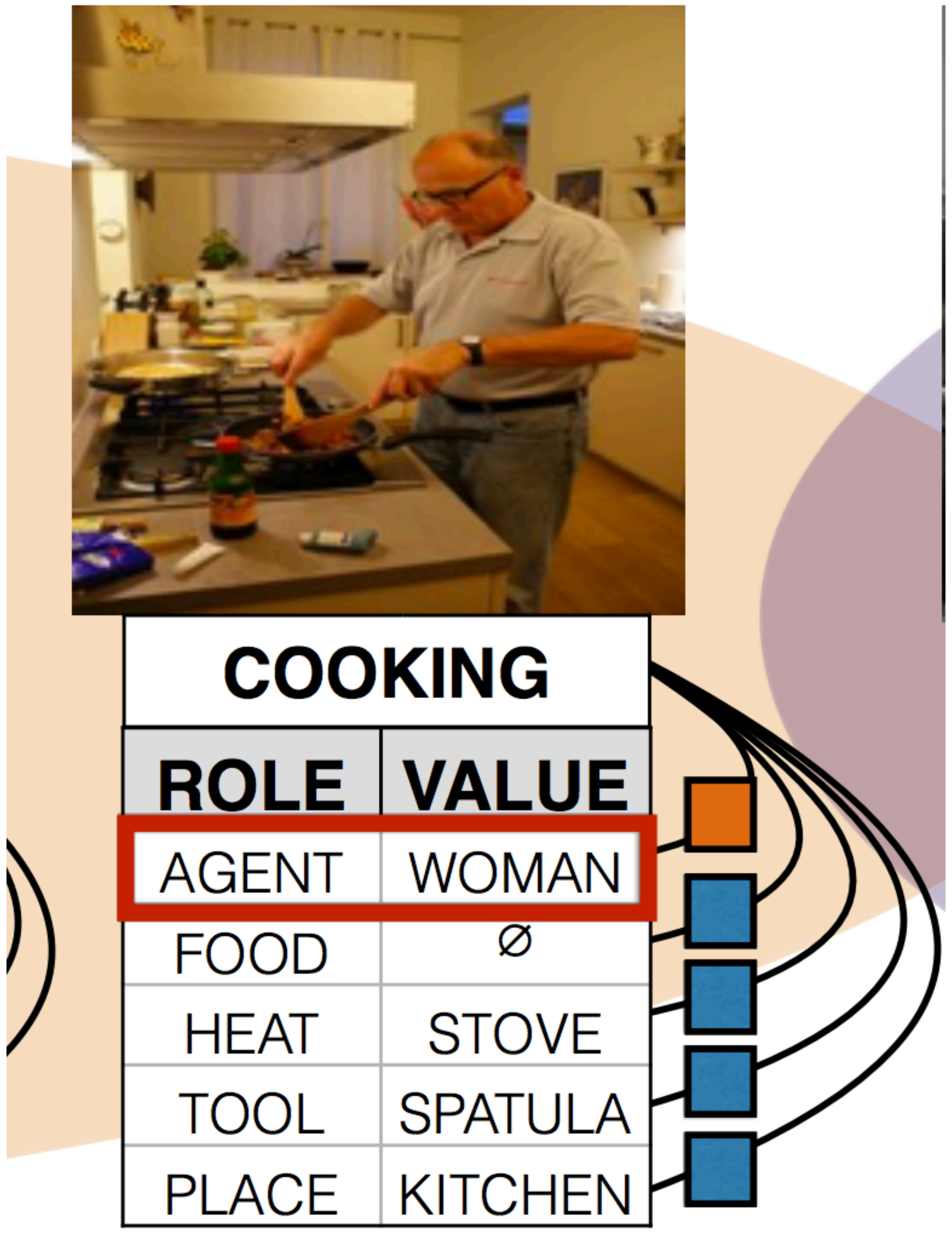
Unethical use: powerful systems can be used for bad ends

Bias Amplification

- Bias in data: 67% of training images involving cooking are women, model predicts 80% women cooking at test time — amplifies bias

- Can we constrain models to avoid this while achieving the same predictive accuracy?

- Place constraints on proportion of predictions that are men vs. women?

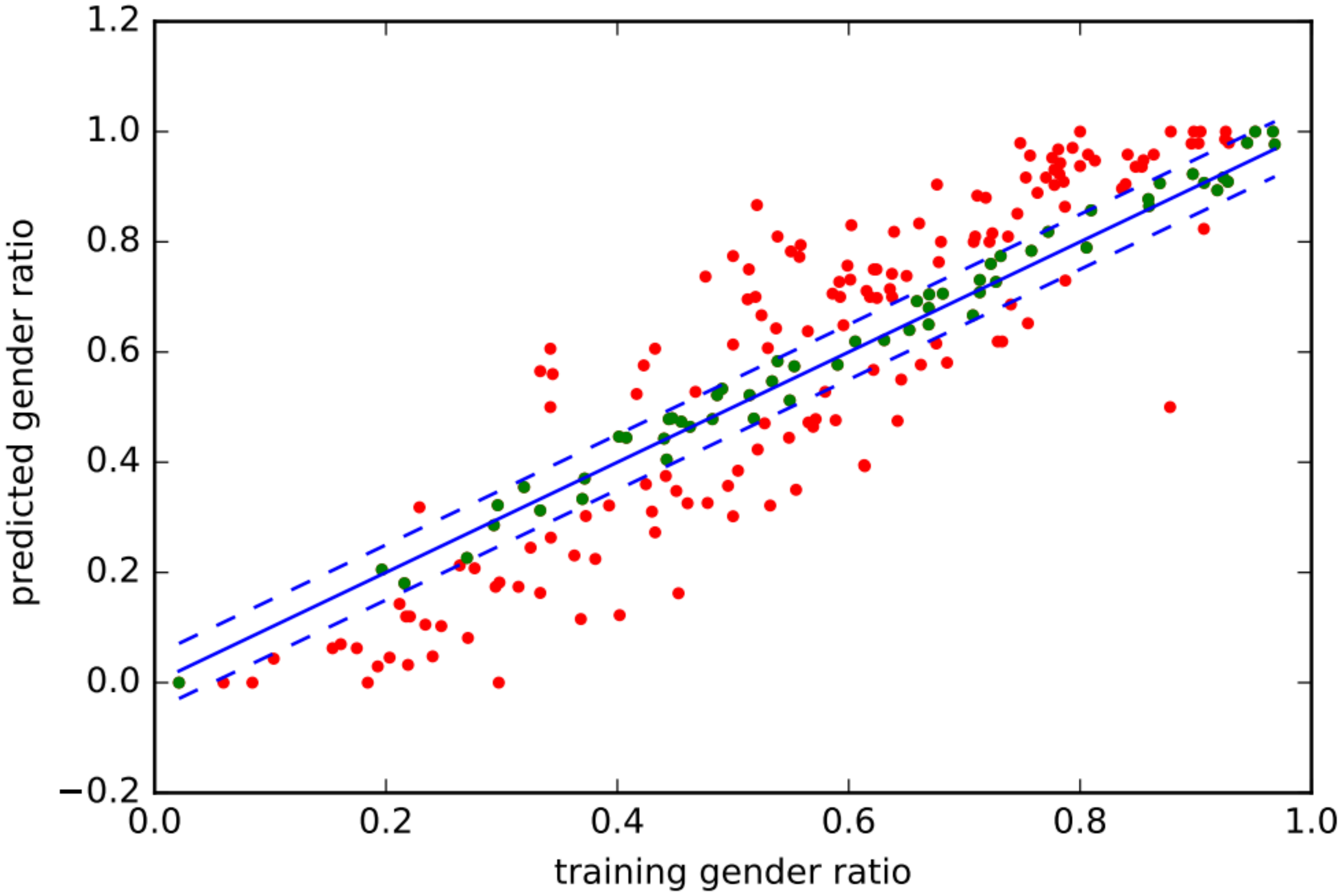


COOKING

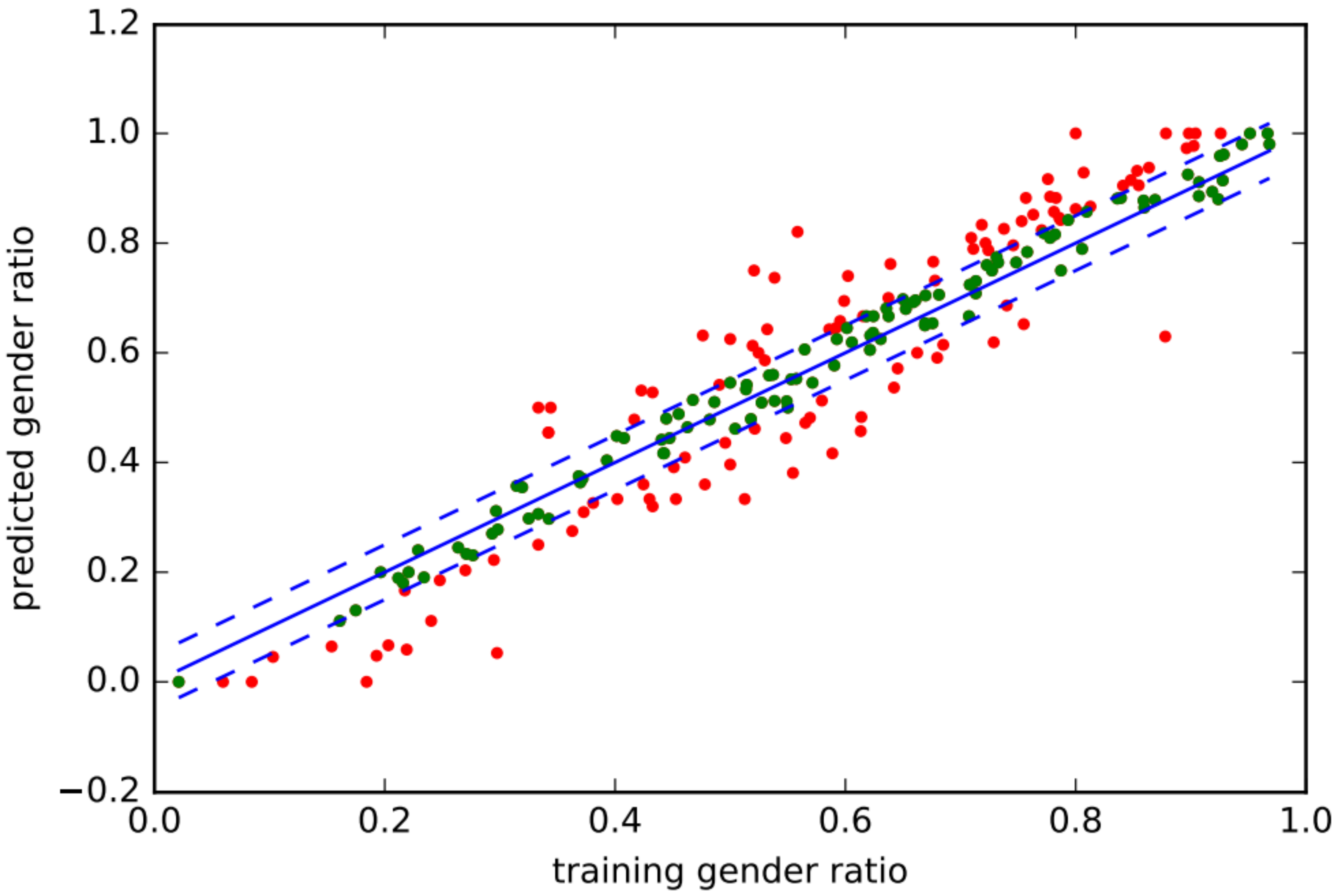
ROLE	VALUE
AGENT	WOMAN
FOOD	∅
HEAT	STOVE
TOOL	SPATULA
PLACE	KITCHEN

Zhao et al. (2017)

Bias Amplification



(a) Bias analysis on imSitu vSRL without RBA



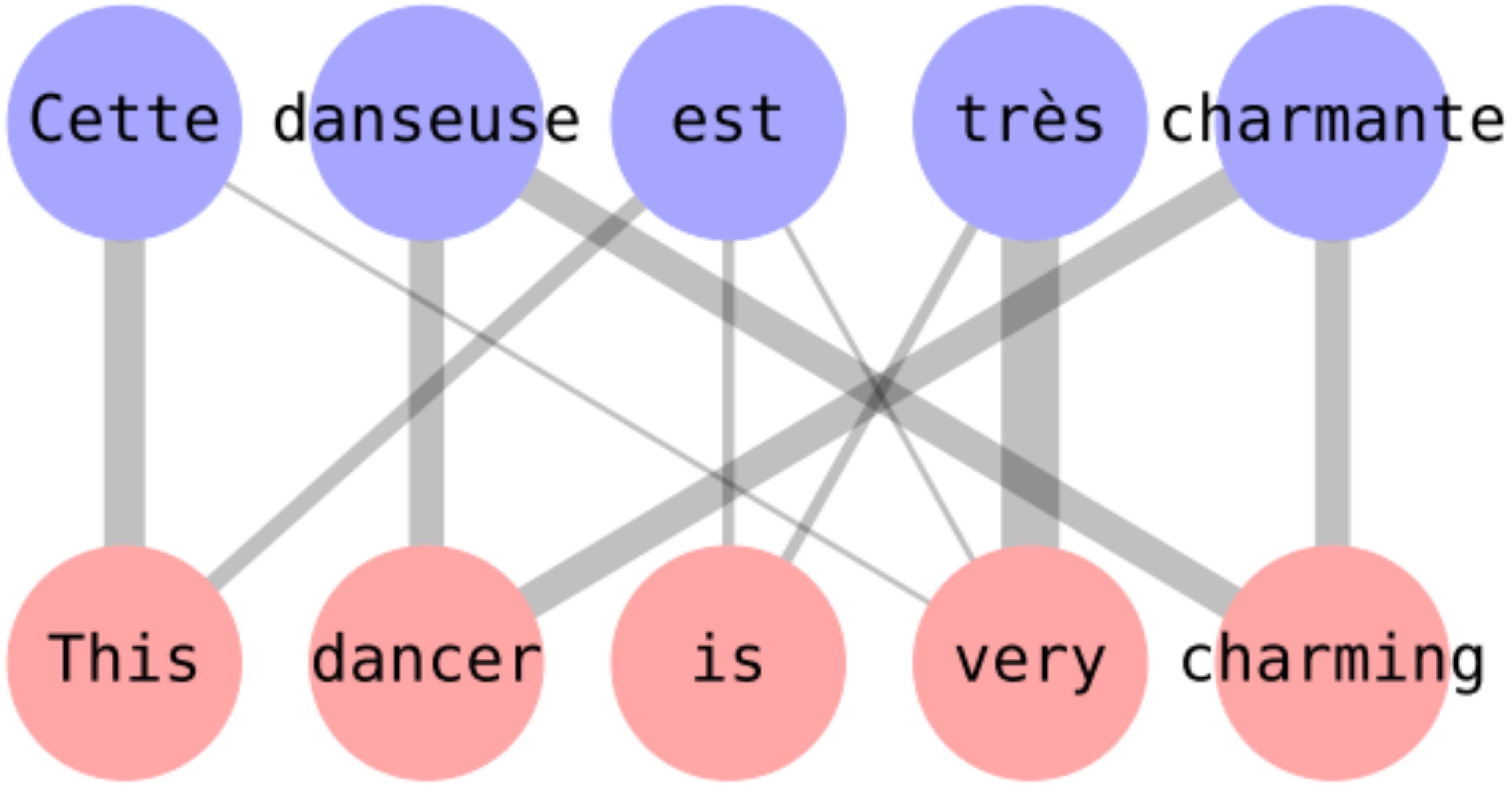
(c) Bias analysis on imSitu vSRL with RBA

Zhao et al. (2017)

Bias Amplification

- English -> French machine translation **requires** inferring gender even when unspecified

- “dancer” is assumed to be female in the context of the word “charming”… but maybe that reflects how language is used?



Alvarez-Melis and Jaakkola (2011)

Bias Amplification: LLMs

- Lots of potential for bias amplification in LLMs and open-ended generation (e.g., reproducing racist jokes at a higher rate than observed in base corpora)

- RLHF does some work to curb this, but lots of ongoing work to make it better

- Other areas of bias amplification: any task involving gender or with gender as a confounder (coreference resolution, parsing someone's occupation)