Research Proposal

1. Research Summary

There are machine learning models to predict if people being released from prison are likely to reoffend (recidivism prediction), but there are important associated questions about ethics. For example the famous case of COMPAS in America where it was considered to be biased against black people (Angwin, 2016), and is still in use today!

The topic of ethics and fairness is important in this area because getting it wrong can have serious implications; unnecessarily depriving someone of their liberty, or releasing someone who is a danger the public. The evidence, however, is that there are serious challenges with providing a fair and ethical model due to the available features and biases in the baseline data.

Considering the challenges with producing a completely fair and unbiased prediction model, methods to highlight potential biases in the output could instead be considered so that due consideration can be taken when interpreting the output. It is proposed to build an ontology based upon the data held about individuals in the criminal justice system data in England and Wales, using metadata from four different sources; magistrate's court, crown court, probation service and prison service (Ministry of Justice, 2020). This would allow:

1. Protected attributes such as race, gender and age (Equality and Human Rights Commission, 2021) to be identified and flagged to either not be used training machine learning models, or used with caution.
2. Additional features not present in the existing datasets to be added to improve fairness and predictability based upon findings from the literature review, such as dynamic factors like residence and education (Andrews & Bonda, 2024; Farrington & West, 1995; Farrington et al., 2017).
3. Actual profiles of protected attributes to be compared with predicted profiles to check for parity, which may or may not equate to fairness.

Note that the proposal is to build an ontology using metadata from Ministry of Justice (2020), not to populate the ontology with actual data because those data are highly sensitive. The principle, and therefore success of the ontology, will be tested outside of the ontology using already anonymised American data (National Institute of Justice, N.D). A statistical analysis will be performed to examine patterns of potential bias in the data to describe how they would be represented in the ontology, but the data itself will not be loaded into the ontology because it has different features.

Research question:

- Can ontology be used to reduce bias when using machine learning to predict recidivism?

Aim:

- Machine learning is in use to predict recidivism but previous studies have indicated ethical issues such as racial bias. This study will show if biases can be identified and mitigated with the use of ontology by creating an ontology of criminal justice in England and Wales, with features identified for safely

predicting recidivism, and features only to be used with caution. Furthermore, the profiles of protected characteristics can be compared with predicted profiles to check for parity.

Objectives:

- Identify features that predict recidivism
- Identify features that potentially introduce unethical biases when predicting recidivism
- Assess if actual recidivism varies between characteristics such as ethnicity, gender and age
- Create an ontology of criminal justice using available metadata and illustrate how the ontology can be used to manage the features to reduce biases, and highlight potential biases in the output to further mitigate risks of bias.

2. Methodology

The research will be quantitative and the method to be used will be “Experiment” (Dawson, 2015: 28).

The first challenge in building a machine learning model to predict recidivism is availability of data. Any ontology designed to assist with the same task therefore faces the same challenge. To make the ontology meaningful in the real world, metadata will be used to design the initial ontology with data already available in the criminal justice system (Ministry of Justice, 2020). Noy & McGuiness (2001) will be referenced alongside analysis of the metadata to create a domain ontology, excluding instances, using four separate metadata sets; crown court, magistrates’ court, prison service and probation service (Ministry of Justice, 2020). Features will be added as informed by the literature review, and any features that are protected characteristics (Equality and Human Rights Commission, 2021) will be highlighted.

Exploratory data analysis (EDA) will be performed using Python in a Jupyter Notebook on anonymised recidivism data from National Institute of Justice (N.D.) to identify correlations between protected attributes such as ethnicity and gender and other attributes, and correlations with recidivism. Hypotheses will be created from the EDA about relationships between the attributes. The hypotheses will be tested for statistical significance using R. Based upon the results, attributes in the ontology will be classified according to their risk of creating unethical biases. Some transformation will be required to approximate similar attributes between the UK metadata and the US actual data. The data won’t actually be loaded into the ontology due to significant differences between the features.

The final ontology will be evaluated, without instances, using an ontology evaluation framework such as Orbst et al (2007) or a method from Raad & Cruz (2015), chosen to best evaluate an empty ontology. The evaluation will be validated against findings of reliable predictors of recidivism from the literature review.

Primary research will not be required.

3. Key literature

The literature review will examine the following topics. Some example references have been included for context but these are far from exhaustive.

- What is recidivism and why and how to predict it? (1,000 words)
    - A common definition of recidivism is “whether a person returns to prison within three years of release” King & Elderbroom (2014: 2)
    - The history of predicting recidivism such as actuarial measures (Borden, 1928; Burgess, 1928)
    - Tools used in different parts of the world such as Statistical Information on Recidivism – Revised (Nafekh & Motiuk, 2002).
    - Predicting with static factors such as criminal history and age vs. dynamic factors such as residence and education (Andrews & Bonda, 2024; Farrington & West, 1995; Farrington et al., 2017).
- How is recidivism predicted in England and Wales? (500 words)
    - The evolution of the Offender Group Reconviction Scale (OGRS) (HM Prison & Probation Service, 2023; Farrington & Davies; 2007; Howard et al., 2009; Moore, 2015).
- How has machine learning been used to predict recidivism to date (real-world and academically)? (1,000 words)
    - Starting with COMPAS (Equivant, 2019) as the most well-known tool (with very public credibility challenges), then moving to explore other machine learning models tested (Curtis, 2018; Kovalchuk et al., 2023; Wang et al., 2010; Zeng et al., 2017), the performance measures used (Ghoneim, 2019), and the implications of using the different measures; more false positives disadvantage the individual, more false negative endanger the public (Caton & Haas, 2020).
- What are the ethical challenges of using machine learning to predict recidivism? Including detailed analysis of the different ethical challenges (they are complex) (3,000 words)
    - Racial discrimination (Angwin et al., 2016; Dieterich et al., 2016; Dressel & Farid, 2021; Thomas, 2023)
    - Age is a strong predictor of recidivism (Bushway & Piehl, 2007; Kleiman et al. 2007; Stevenson & Slobogin, 2018).
    - There are demonstrable differences in recidivism between male and female offenders (Skeem et al., 2015), so is it OK to use this protected characteristic (Skeem & Lowenkamp, 2020)
    - The importance of explainability to ethical fairness (Babad & Chun, 2023; Rodu & Baiocchi, 2023; Walmsley, 2021)
    - Reducing racial bias reduces accuracy (Skeem & Lowenkamp, 2020) because the baseline dataset is already biased (Thomas, 2023).
- How have ontologies been used to address ethical challenges (500 words)

- Short section to say where ontologies have been used (Franklin et al., 2023; Kasirzadeh & Smart, 2021; Lehnert, 2021), but nothing like proposed in this dissertation, hence the dissertation is new and innovative.

4. Human Participants

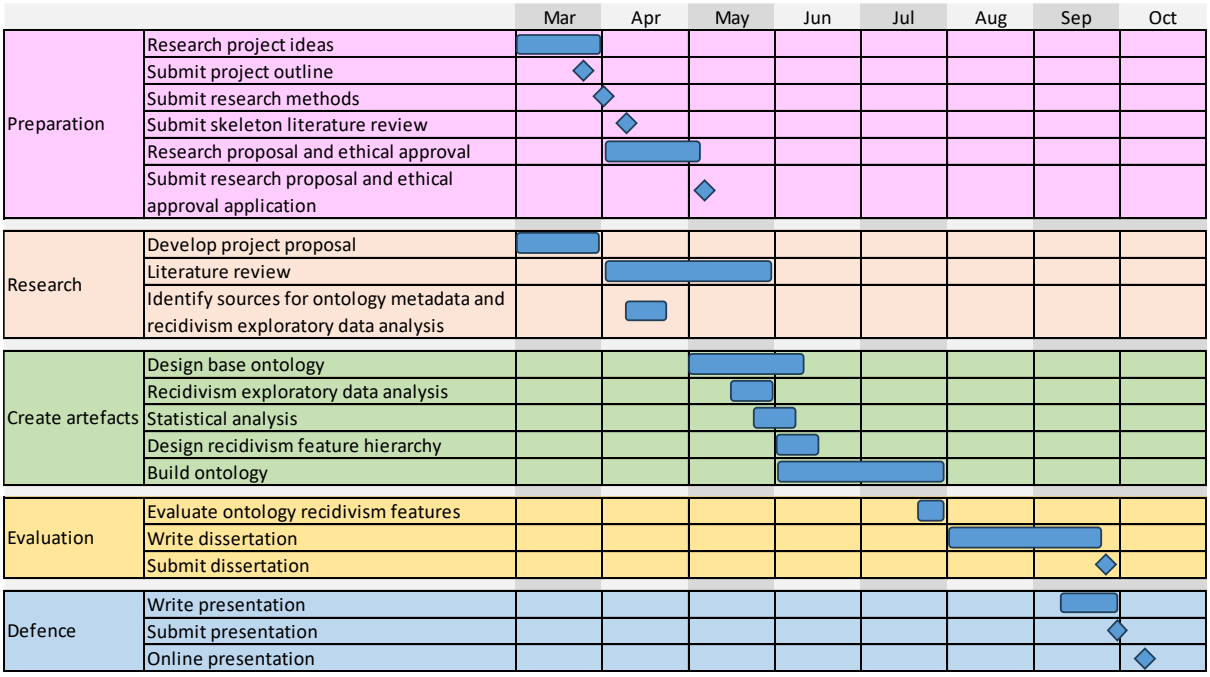
There will be no human participants because primary research will not be required.

5. Timeline

Artefacts that will be created are:

- Exploratory data analysis (EDA) of a non-UK data set (National Institute of Justice, N.D.) (UK not available) to identify features that correlate with protected features such as ethnicity and gender.
- Statistical analysis of relevant feature relationships. For example:
  - Is there a statistically significant difference between actual recidivism rates for different ethnicities, genders or ages?
  - Are the correlations identified in the EDA statistically significant?
- Ontology of criminal justice using metadata from magistrate court, crown court, prison service and probation in England and Wales (Ministry of Justice, 2020) as a baseline, supplemented with additional features identified in the literature review, if applicable. The features will be organised to provide a method for segregating data using the indicative (non-UK) findings from the EDA and statistical analysis to show differences between the classes in both baseline data and predictions. For example, is there a difference in the recidivism predictions between black and white offenders? Does the same difference exist in the base data? From there a risk assessment can be made on how “safe” the prediction profile is (not an individual prediction). This will be using metadata only, so it won’t actually show the predictions, just how they would be presented.

A project plan is shown below. Note that in order to complete in time this assumes the literature review is already underway, which is the case. In fact the research is complete and the write-up is almost complete, as summarised above.



Key risks:

- Time to get it all done alongside a full time job! Sticking to the project plan will be crucial, hence significant progress to date on the literature review.
- There is a risk that the exploratory data analysis and statistical analysis don't uncover features that can be used to address the ethical challenges. In that case the resultant ontology and conclusions will be modified to visualising results in future to uncover biases rather than addressing previously identified biases.

References

Andrews, D.A. & Bonta, J. (2024) *The psychology of criminal conduct.* 7th ed. New York: Routledge.

Angwin, J., Larson, J., Mattu, S. & Kirchner, L. (2022) 'Machine Bias', in: Martin, K. (ed) *Ethics of Data and Analytics*. New York: Auerbach Publications. 254-264.

Borden, H.G. (1928) Factors for predicting parole success. *Journal of the American Institute of Criminal Law and Criminology*, 19(3): 328-336.

Burgess, E.W. (1928) 'Factors determining success or failure on parole', in: Bruce, A., Burgess, E. & Harno, A. (eds) *The workings of the indeterminate sentence law and the parole system in Illinois*. Springfield, IL.: Illinois State Board of Parole. 221–234

Bushway, S.D. & Piehl, A.M. (2007) The inextricable link between age and criminal history in sentencing. *Crime & Delinquency*, 53(1): 156-183. https://doi.org/10.1177/0011128706294444

Caton, S. & Haas, C. (2020) Fairness in machine learning: A survey. *ACM Computing Surveys*, 56(7): 1-38. https://doi.org/10.1145/3616865

Curtis, J. (2018) On using machine learning to predict recidivism. Ph.D. thesis, Texas Tech University. Available from: https://ttu-ir.tdl.org/server/api/core/bitstreams/8e745777-200b-45d2-94a1-84355598d2ba/content [Accessed 22 April 2024].

Dieterich, W., Mendoza, C. & Brennan, T. (2016) *COMPAS risk scales: Demonstrating accuracy equity and predictive parity*. Northpointe Inc. Research Department.

Dressel, J. & Farid, H. (2021) The dangers of risk prediction in the criminal justice system. *MIT Case Studies in Social and Ethical Responsibilities of Computing*. https://doi.org/10.21428/2c646de5.f5896f9f

Equality and Human Rights Commission (2021) Protected characteristics. Available from: https://www.equalityhumanrights.com/equality/equality-act-2010/protected-characteristics [Accessed 27 April 2024].

Farrington, D.P. & Davies, D.T. (2007) Repeated contacts with the criminal justice system and offender outcomes. Statistics Canada. Available from: https://www.crim.cam.ac.uk/sites/www.crim.cam.ac.uk/files/statcanf.pdf [Accessed 25 April 2024].

Farrington, D.P. & West, D. J. (1995) 'Effects of marriage, separation and children on offending by adult males'. in Hagan, J, (ed) *Current Perspectives on Aging and the Life Cycle. vol. 4: Delinquency and Disrepute in the Life Course*. Greenwich, CT: JAI Press. 249-281.

Farrington, D.P., Gallagher, B., Morley, L., Ledger, R.J.S. & West, D.J. (2017) 'Unemployment, school leaving, and crime'. in Farrall, S. (ed) *The Termination of Criminal Careers*. London: Routledge. 101-122.

Franklin, J.S., Powers, H., Erickson, J.S., McCusker, J., McGuinness, D.L. & Bennett, K.P. (2023) 'An Ontology for Reasoning About Fairness in Regression and Machine Learning', *Fifth Iberoamerican and the Fourth Indo-American Knowledge*

Graphs and Semantic Web Conference. University of Zaragoza, Zaragoza, Spain, 13-15 November. Cham, Switzerland: Springer Nature. 243-261.

Ghoneim, S. (2019) Accuracy, Recall, Precision, F-Score & Specificity, which to optimize on? Available from: https://towardsdatascience.com/accuracy-recall- Page 15 precision-f-score-specificity-which-to-optimize-on-867d3f11124 [Accessed 22 April 2022].

HM Prison & Probation Service (2023) Risk of Serious Harm Guidance 2020 v3. London: HM Prison & Probation Service. Available from: https://assets.publishing.service.gov.uk/media/652cf8c9697260000dccf834/Risk_of_Serious_Harm_Guidance_v3.pdf [Accessed 17 April 2024].

Howard, P., Francis, B., Soothill, K. & Humphreys, L. (2009) OGRS 3: The revised offender group reconviction scale. London: Ministry of Justice.

Kasirzadeh, A. & Smart, A. (2021) March. 'The use and misuse of counterfactuals in ethical machine learning', *ACM Conference on Fairness, Accountability, and Transparency*. Virtual Event, Canada, 3-10 March. New York: Association for Computing Machinery. 228-236. https://doi.org/10.1145/3442188.3445886

King, R.S. & Elderbroom, B. (2014) *Improving recidivism as a performance measure*. Washington, DC: Urban Institute. Available from: http://hint-magazine.com/wp-content/uploads/2014/10/413247-improving-recidivism.pdf [Accessed 01 April 2024].

Kleiman, M., Ostrom, B.J. & Cheesman, F.L. (2007) Using risk assessment to inform sentencing decisions for nonviolent offenders in Virginia. *Crime & Delinquency*, 53(1): 106-132. https://doi.org/10.1177/0011128706294442

Kovalchuk, O., Karpinski, M., Banakh, S., Kasianchuk, M., Shevchuk, R. & Zagorodna, N. (2023) Prediction machine learning models on propensity convicts to criminal recidivism. *Information*, 14(3): 1-15. https://doi.org/10.3390/info14030161

Lehnert, A. (2021) Ontologies and Ethical AI. Available from: https://www.synaptica.com/ontologies-and-ethical-ai/ [Accessed 02 May 2024].

Ministry of Justice (2020) Ministry of Justice: Data First. Available from: https://www.gov.uk/guidance/ministry-of-justice-data-first [Accessed 18 March 2024].

Moore, R. (2015) A compendium of research and analysis on the Offender Assessment System (OASys). London: National Offender Management Service. Available from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/449357/research-analysis-offender-assessment-system.pdf [Accessed 18 April 2024].

Nafekh, M. & Motiuk, L.L. (2002) *The statistical information on recidivism, revised 1 (SIR-R1) scale: a psychometric examination*. Ottawa, Ontario: Correctional Service of Canada, Research Branch.

National Institute of Justice (N.D.) Recidivism Forecasting Challenge. Available from: https://nij.ojp.gov/funding/recidivism-forecasting-challenge [Accessed 18 March 2024].

Noy, N.F. & McGuinness, D.L. (2001) *Ontology Development 101: A Guide to Creating Your First Ontology*. Knowledge Systems Laboratory.

Obrst, L., Ceusters, W., Mani, I., Ray, S. & Smith, B. (2007) The evaluation of ontologies: Toward improved semantic interoperability. *Semantic web: Revolutionizing knowledge discovery in the life sciences*: 139-158.

Raad, J. & Cruz, C. (2015) 'A survey on ontology evaluation methods', *International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*. Lisbon, Portugal, 12-14 November 12-14. Cham, Springer. 179-186. https://doi.org/10.5220/0005591001790186

Skeem, J. & Lowenkamp, C. (2020) Using algorithms to address trade-offs inherent in predicting recidivism. *Behavioral Sciences & the Law*, 38(3): 259-278. https://doi.org/10.1002/bsl.2465

Stevenson, M.T. & Slobogin, C. (2018) Algorithmic risk assessments and the double-edged sword of youth. *Behavioral sciences & the law*, 36(5): 638-656. https://doi.org/10.1002/bsl.2384

Thomas, S. (2023) *The Fairness Fallacy: Northpointe and the COMPAS Recidivism Prediction Algorithm*. Ph.D. thesis, Columbia University. Available from: https://academiccommons.columbia.edu/doi/10.7916/ab13-jf83 [Accessed 29 April 2024].

Wang, P., Mathieu, R., Ke, J. & Cai, H.J. (2010) 'Predicting criminal recidivism with support vector machine'. International Conference on Management and Service Science. Wuhan, China, 24-26 August. IEEE. 1-9. https://doi.org/10.1109/ICMSS.2010.5575352

Zeng, J., Ustun, B. & Rudin, C. (2017) Interpretable classification models for recidivism prediction. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 180(3): 689-722. https://doi.org/10.1111/rssa.12227