BACHELOR’S THESIS EXAMINER REPORT

PPE – Bachelor’s in Politics, Philosophy and Economics
Faculty of Social Sciences, Charles University

Thesis title:	Artificial Intelligence as a Challenge to Social Justice in the Light of the Theories of J. Rawls and M. Walzer
Student’s name:	Johanna Reichart
Referee’s name:	Petr Špecián

Criteria	Definition	Maximum	Points
Major Criteria			
	Contribution and argument (quality of research and analysis, originality)	**50**	45
	Research question (definition of objectives, plausibility of hypotheses)	**15**	13
	Theoretical framework (methods relevant to the research question)	**15**	13
Total		***80***	70
Minor Criteria			
	Sources, literature	**10**	9
	Presentation (language, style, cohesion)	**5**	5
	Manuscript form (structure, logical coherence, layout, tables, figures)	**5**	3
Total		***20***	17
TOTAL		***100***	88

Plagiarism-check (URKUND) match score: 13 %
[NB:] If the plagiarism-check (URKUND) match score is above 15%, the reviewer has to include his/her assessment of the originality of the reviewed thesis in his/her review.

Reviewer’s commentary according to the above criteria (min. 1800 characters including spaces when recommending a passing grade, min. 2500 characters including spaces when recommending a failing grade):

Johanna Reichart's thesis, “Artificial Intelligence as a Challenge to Social Justice in the Light of the Theories of J. Rawls and M. Walzer,” presents a bold exploration of a significant topic in the intersection of technology, philosophy, and social justice. The work demonstrates the author’s high level of analytical, philosophical, and writing competence. Nonetheless, while the thesis makes an interesting contribution to the field, there are some limitations that warrant discussion.

Strengths:

1. The thesis engages with a broad range of academic sources, demonstrating the author’s excellent grasp of complex literature and ability to synthesize diverse perspectives.
2. The approach to the topic is ambitious, reflecting a clear effort to make a substantial contribution to the debate rather than merely fulfilling degree requirements.
3. The writing is of high quality, with excellent command of English, making the thesis very readable and accessible.

4. The research question is well-formulated (p. 8), and the author makes systematic and well-organized efforts to answer it throughout the thesis.
5. The work presents several courageous, original ideas. For instance, the suggestion that "non-liberal, hierarchical peoples are much more likely to influence well-ordered peoples rather than vice versa" (p. 32) is thought-provoking,), although I am not quite convinced by the underlying argument since the key AI developments are taking place in the US with China struggling behind.

Weaknesses:

1. **The definitions of AI employed in the thesis are problematic and not well-chosen**:
   i) The primary definition, "AI as a digital tool that can develop human-like, autonomous behaviours from data input" (p. 8), is both vague and presumably too narrow. It may fail to encompass narrow instantiations of AI, such as social media algorithms, which play an important role in the thesis, and not particularly "human-like" or "autonomous" in any intuitive sense of these terms. The thesis does not provide its own clear definitions of what constitutes "human-like" or "autonomous" behavior, leaving these concepts open to interpretation.

   ii) The distinction between 'strong AI' and 'weak AI', borrowed from Hermansyah et al. (2023), creates further confusion: Boundary between "less" and "more" cognitively complex tasks, which is crucial for this distinction to be meaningful, is not even vaguely established.
2. **The thesis attempts to cover too much ground**, often resulting in arguments that are too abstract. The wide-ranging definition of AI encompasses disparate technologies with varying social impacts—social media algorithms, large language models, various classifiers including facial recognition etc. (presumably, narrowly specialized AIs, such as Google's AlphaFold would also qualify)—making it challenging to draw meaningful conclusions that would apply universally. Similar to much philosophical literature in the field, the thesis does not always succeed in doing that.
3. **The thesis suffers from a lack of specific examples** of AI-related problems:
   i) The only substantial example provided is China's Social Credit System (pp. 16-17), and even this is not explored in sufficient depth. This scarcity of concrete examples forces the thesis to remain at an overly abstract level of argument, limiting the cogency and persuasiveness of the claims being made.

   ii) For instance, let us consider the thesis's treatment of AI in education: "algorithmic biases would reinforce existing injustices and create new ones (Filgueiras, 2023, p. 7) which, in this case, would mean that not all are assessed equally. Some individuals would subsequently be provided with fewer opportunities to learn and be deprived of opportunities to educate themselves further." However, this seems as a one-sided view, focusing only on potential negative outcomes without considering possible benefits. AI, particularly in the form of large language models, could democratize access to information and learning resources on an unprecedented scale. Also, while algorithmic biases are a genuine concern, their effects are not predetermined and may be mitigated through careful design and implementation.
4. Some of the author's **views on political processes and state functions appear idealized**. For instance, the characterization of the EU's AI Act as "a prime example of how to regulate AI in line with Rawls' theory of justice" (p. 30) overlooks the complexities and potential shortcomings of real-world political processes with their democratic deficits, pressure group maneuvers, politicking, etc.

5. **Some factual claims about AI capabilities are outdated or inaccurate**, particularly those relying on older sources like Brożek & Janik (2019). This leads to some misconceptions about current AI capabilities, which could affect the philosophical conclusions drawn. For instance, large language models do undergo a process of "moral learning" (p. 36) via RLHF, also they do posses some of the human propensities to irrationality and do have significant emotional intelligence.
6. Formal note: The conclusion of the thesis is too brief, and the numbering of subchapters is excessively deep and potentially confusing for readers.

Overall Assessment:

The thesis represents a high-quality effort to grapple with complex issues at the intersection of artificial intelligence and social justice theory. The author demonstrates strong analytical skills, a broad understanding of relevant literature, and the ability to formulate original ideas. However, the thesis would benefit from a more focused scope, more concrete examples, and updated information on AI capabilities. These weaknesses prevent me from proposing the highest possible mark. However, a persuasive defense might perhaps still push the thesis toward an A.

Proposed grade (A-B-C-D-E-F): B

Suggested questions for the defence are:

You claim that "an individual who is brought up in an environment shaped by mis- and disinformation, in which one is hardly confronted with disagreement, will find it hard to rationally deliberate about principles of justice considering other persons' perspectives." To what extent are we living in such an environment? For instance, would you say that social media shield one from disagreement? These platforms are often painted as outrage machines continuously triggering people by showing them discordant views. Also, there is a lively debate about online cancellations. Please, discuss.

Is 'AI' a philosophically productive concept, given the growing diversity of technologies hidden under this label? Can you defend your definition of AI or propose some alternative definition that could have been more philosophically productive (with the benefit of hindsight)?

I recommend the thesis for final defence.

Ing. Petr Špecián, Ph.D.

Digitálně podepsal Ing. Petr Špecián, Ph.D.
Datum: 2024.08.26 09:15:27 +02'00'

Referee Signature

Overall grading scheme at FSV UK:

TOTAL POINTS	GRADE	Quality standard
91 – 100	**A**	= outstanding (high honor)
81 – 90	**B**	= superior (honor)
71 – 80	**C**	= good
61 – 70	**D**	= satisfactory
51 – 60	**E**	= low pass at a margin of failure
0 – 50	**F**	= failing. The thesis is not recommended for defence.