Handbook

for

3rd Practical Philosophy Graduate Student and Postdoctoral Researcher Conference

Society for Practical Philosophy Turkey

9-10 December 2022

Conference Program: (GMT +3)

Friday, December 9th:

16:50-18:00 Keynote Speaker: Benjamin Ferguson
"Exploitation's Grounding Problem"

18:00-18:50 "Kant's (Integrated) Model of Moral Judgment"
by Stefano Pinzan

coffee/dinner break

19:20-20:10 "Vocation and the Williamsian Problem"
by Barnabas Agota

20:10-21:00 "Helping in Solidarity with a Movement: The Practical Importance of a Hybrid Approach to the Ethics of Collective Action"
by Maddie Youngman

Saturday, December 10th:

16:00-17:00 Keynote Speaker: Helga Varden
"Some Kantian Thoughts on Method: Transforming the Social Contract Tradition's Distinction between the State of Nature and Civil Society"

coffee break

17:20-18:10 "Content moderation in Social Media Platforms"
by Paride Del Grosso

18:10-19:00 "Overcoming the Particularity Problem"
by Avontay Williams

coffee/dinner break

19:20-20:10 "AlphaGo, Intentionality, and the Prospect of Artificial Moral Patiency"
by Tuğba Yoldaş

20:10-21:00 "Epistemic Hierarchy"
by Idowu Odeyemi

Exploitation's Grounding Problem

Benjamin Ferguson

Professor of Philosophy, University of Warwick

https://www.benjaminferguson.org/

Abstract: Standard accounts of what makes exploitation wrong grounds its wrong, at least in part, in distributive unfairness. That is, when A exploits B he wrongs her by taking a greater share of the benefits from their interaction than he ought. I argue that this standard account does not succeed; exploitation's wrong cannot be grounded on distributive unfairness. I assume that distributive unfairness is pro tanto wrong. I then show that many cases of exploitation that also involve distributive unfairness are actually wrong because the exploited party fails to provide morally valid consent. However, I argue that in situations where consent is morally valid, this it is also morally transformative and overrides distributive unfairness's pro tanto wrong. Consequently, exploitations are either wrong because they lack morally valid consent, or they are not all things considered wrong.

Some Kantian Thoughts on Method:

Transforming the Social Contract Tradition's Distinction between the State of Nature and Civil Society

Helga Varden

Professor of Philosophy, University of Illinois at Urbana-Champaign

https://helgavarden.com/

Abstract: (Early) Modern social contract theories were an important human invention because they envision justice in terms of the protection and realization of each person's freedom and equality. These theories characteristically reject the idea that legal and political institutions are grounded in an alleged natural ordering or hierarchy of human beings, and instead argue that only government by a public (and not private) authority could fulfil the idea of justice as freedom and equality for all. To be authoritative and not just powerful, governing institutions must be shared as ours in this irreducible sense. I first outline how Kant's ideal account of rightful freedom brilliantly transforms the tradition as found in the works of Hobbes, Locke, and Rousseau, before proposing a way to see Kant's two-layered non-ideal theory—his accounts of human nature ("moral anthropology") and historical societies ("the principle of politics")—as complementing his ideal theory of rightful freedom. This enables us to envision a conception not only rightful external freedom but rightful human freedom in particular societies—with their histories—on planet Earth. With these arguments at hand, we can also better appreciate the importance of realizing that the four possible political conditions for Kant—anarchy, barbarism, despotism, and republic—are ideas of reason, which means that they are never perfectly realized. Hence, historical societies we are not either in the state of nature or in civil society. In addition, in historical societies founded on principles of freedom, there are pockets of injustice or pockets that are devoid of justice that can only be captured by means of one of the three political ideas that are not constitutive of the republican legal-political framework. Kant's four ideas therefore give us more tools with which to capture the nature of different political forces and challenges facing us in our historical societies.

Kant's (Integrated) Model of Moral Judgment

Stefano Pinzan

Vita-Salute San Raffaele University (Milan)

Abstract: Data from neuroscientific studies often determine conditions of psychological feasibility (Kauppinen 2014) for normative theories, with the risk that many of the latter will be deemed unfeasible. Kant's moral philosophy seems to lie within the group of theories exposed to risk, especially because of the role that emotions play in moral judgment. Various studies have, in fact, shown that, as subjects formulate moral judgments, brain areas connected to emotionality are activated (Green 2014). Consequently, these studies support the need for at least an integrated model of moral judgment, with essential roles for both the emotional and the rational components (De Caro, Marraffa 2016); a model that the German philosopher's theory seems unable to support. The Kantian agent is indeed often portrayed as one who must judge and act without taking emotional life into account, but only through pure rationality.

Furthermore, Kant's theory must respond not only to the challenges presented by neuroscience, but also to the more general renewed interest in the anthropological-emotional structure of the subject (Pulcini 2020) that has arisen from the revaluation of sentimentalist thought; that body of thought is considered better suited to emphasizing the nuances of emotional life and the normative role of emotions.

In this paper I argue for a different view. Indeed, I aim to clarify the real relation between reason and emotions within Kant's ethics with the aim of showing how the German philosopher ultimately offers an integrated model of moral judgment that does not deviate from contemporary empirically informed models.

First, I argue that the traditional portrayal of Kant's view on emotions is based on an incomplete reading of his works. Kant is indeed fully aware of the sensible and emotional dimension of moral experience. This is already clear from his analysis of respect, a crucial node of his ethics, as the expression of the link between the principle of morality and the need for a sentimental recognition. Moreover, Kant offers an elaborated analysis of emotional life. He distinguishes between affects, passions, and feelings and assesses the different ways in which they interact with practical reason. The moral agent must act out of duty, but this does not entail the extirpation of her own sensibility. On the contrary, some "pathological" feelings, as the sympathetic ones, may even play a positive, and morally worthy, role: if properly cultivated (Cohen 2018), unlike affects that must be disciplined (Eran 2020), such feelings could participate in the construction of good character (Felicitas Munzel 1999, Pinzan 2022a).

In particular, it is possible to argue that emotions play an active role within the process of moral judgment in Kant. This process can be divided into two fundamental moments: the determination of the content of the maxim and the critical scrutiny of the form of the maxim. Emotions participate in the construction of the maxim. The latter is for Kant the place of the particular (Pinzan 2022b), the subjective determination of the will: the faculty of desire and the emotions participate in the constitution of its object, orienting and guiding the subject, as explained by Nancy Sherman with her "perceptual claim" (Sherman 1990). It could be argued that in this first moment, namely that of the construction of the maxim, emotions have a normative value and that the maxim itself can be configured as a form of prima facie moral judgment. At the same time, it would be better to speak of a material normativity of emotions, that is a non-moral form of normativity, since morality for Kant is not played out on the level of matter, but rather on the level of form. For this reason, the maxim, which we have also characterized as a form of prima facie moral judgment, must be subjected to critical scrutiny by practical reason through recourse to the categorical imperative, which, unlike the emotions, has a morally normative value. Such a critical/reflective passage makes it possible to move from a subjective dimension of judgment to one that is universally valid, intelligible, and shareable among agents. However, such a reinterpretation of Kant's model of moral judgment diminishes the distance from contemporary integrated models: there is, in fact, an important role for the emotional component of the subject, which guides the subject in the first phase of the judgment process. Thereafter, the reflexive component of the agent acts as a means of checking and balancing the material processed up to that point. In particular, Kant's model summarised in this way can come close to the model of educated intuitions offered by Sauer.

Thus, not only is Kant able to respond positively to the challenges posed by neuroscience in the form of conditions of psychological feasibility of his theory, but he also offers an integrated model of moral judgment along the lines of contemporary empirically informed models. For this reason, Kant's ethics and its specific normative insights must still be a point of reference today within the debate on moral judgment and, more generally, on the interplay between emotions and reason.

References

Cohen A. (2018), "Kant on Moral Feelings, Moral Desires and the Cultivation of Virtue", *International Yearbook of German Idealism*, De Gruyter, Berlin, pp. 3-18.

De Caro M., Marraffa M. (2016), "Debunking the Pyramidal Mind: A Plea for Synergy Between Reason and Emotions", *The Journal of Comparative Neurology*, n.524, 1695-1698.

Eran U. (2020), "Which Emotions Should Kantians Cultivate (and Which Ones Should they Discipline)?", *Kantian Review*, v.25, n.1, 53-76.

Felicitas Munzel G. (1999), *Kant's Conception of Moral Character*, UCP, Chicago.

Greene J.D. (2014), "Beyond Point-and-Shoot Morality: Why Cognitive (Neuro)Science Matters for Ethics", *Ethics*, vol. 124, n. 4, 695-726.

Kauppinen A. (2014), "Ethics and Empirical Psychology – Critical Remarks to Empirically Informed Ethics", M. Christen, C. van Schaik, J. Fischer, M. Huppenbauer, and C. Tanner (ed.), *Empirically Informed Ethics: Morality Between Facts and Norms*, Springer, London.

Pinzan S. (2022b), "A Kantian Ethics of Care?", *Notizie di Politeia*, XXXVIII, n. 147, pp. 24-38.

Pinzan S. (2022a), "Freedom and Sensibility in Kant: An Anthropological-Psychological View", *Filosofia Morale/Moral Philosophy*, vol. 1, pp. 63-83.

Pulcini E. (2020), *Tra cura e giustizia. Le passioni come risorsa sociale*, Bollati Boringhieri, Torino.

Sherman N. (1990), "The Place of Emotion in Kantian Morality", in O. Flanagan and A.O. Rorty (eds.), *Identity, Character, and Morality*, MIT Press, Cambridge, pp.151-170.

Vocation and the Williamsian Problem

Barnabas Agota

Eötvös Loránd University, Budapest

Abstract: One pressing question in the field of normative ethics is the question of partial and impartial morality. As I see, what is at stake at this debate is our deep personal affections and preferences. How can we follow them when they clash with morality? If we can not answer this question we have to surrender to morality, and that is a problem, because it seems like we do have a moral right to follow our deep personal desires and it is also necessary because these personal projects are the things that give meaning to our life. As the great British philosopher, Bernard Williams formulated this problem: ethical theories hurt the agent's integrity. I will call this problem as the Williamsian problem.

I suppose that the question can be answered through the concept of vocation, because it seems to be a concept that combines the unconditionality of morality and the important role of individual preferences in the conduct of life. After outlining the Williamsian problem, I will introduce the concept of vocation, its possible conceptions, and the advantages and disadvantages of each concept. Since the notion of vocation first appears in Judeo-Christian culture, I begin my analysis with the theistic account of vocation, then I present the general characteristics of the notion of vocation, and turn to secular theories. These include self-actualization theory, in which vocation is the fulfillment of one's true self, and personal choice theories, in which vocation is the result of the autonomous agent's choice.

As it turns out, the biggest question is what gives the normative power to these distinct concepts of vocation. If we want to say that it is morally justified in some cases to follow the demand of one's vocation instead of the demands of general ethical theories we need to answer the question why vocation has bigger normative power than the general ethical code.

It is necessary to ground this normative power on something. The theistic account of vocation has a good answer to that, because in this outlook vocation is a call from God. It seems plausible that the word of God is stronger than the general demands of ethics. The problem with this account is that it is too expensive, one needs to accept theism in order to accept the theistic account of vocation. The secularized accounts of vocation do not perform very well in this regard, I argue, it is difficult to attribute such strong normative power to naturalized accounts of vocation that is stronger than normative ethical theories.

Self-actualization and autonomous personal choice are valuable indeed, but do not seem to be stronger than general ethics. Therefore, I think, the most

exciting question is how can we naturalize the notion of vocation which means building a plausible and normatively strong, secularized account of vocation. At the last part of the talk, I intend to introduce my attempt to formulate such a view. This is a self-actualization account that claims that our true self is not as instructive as it might seem like for the first glance. To actualize it we need further orientation and I argue that existing social practices and traditions are good means for that.

Helping in Solidarity with a Movement: The Practical Importance of a Hybrid Approach to the Ethics of Collective Action

Maddie Youngman

University of Alberta

Abstract: In *collective action cases*, sufficient individual contributions will together cause a beneficial outcome, but each individual contribution seems unlikely to make a difference to the outcome. There seem to be strong moral reasons to contribute in these cases, but it can seem that there is no justification for contributing if our acts are superfluous to the outcome. Two promising approaches to explaining why we should contribute in these cases are the expected consequences approach of authors like Shelly Kagan (2011) and the helping-based approach developed by Julia Nefsky (2016). The former suggests that in collective action cases our individual acts do in fact have a small chance of making a difference to the outcome occurring, which explains why we should contribute. Nefsky, by contrast, argues that in collective action cases our individual contributions can non-superfluously *help* to cause the outcome *without* making a difference to whether it occurs, and that our reasons to contribute are reasons to help bring the outcome about.

In this paper, I argue that the best account of our reasons to contribute in collective action cases combines elements of both of these approaches. This *hybrid account* agrees with Nefsky's helping-based approach that our reasons to contribute are reasons to help bring the beneficial outcome about, which we can do even if we fail to make a difference to its occurrence. But the hybrid account holds that one's act cannot actually help if it is certain to make no difference, and that the degree to which it can be expected to help, and our reasons to try to help by performing it, are proportional to our chances of making a difference and the degree to which the outcome would be beneficial. It thus agrees in principle with the expected consequences approach about the strengths of our reasons to contribute in various ways to various collectively caused outcomes.

I contend that this hybrid account has important theoretical and practical advantages. It shares with Nefsky's helping-based account the advantage of giving a more inherently compelling explanation of why we should contribute in collective action cases where our acts have remote chances of making a difference to extremely beneficial outcomes. In these cases, it is much easier to see the point of doing what has a significant chance of helping, if only to a very small degree, to bring about an extremely beneficial outcome, than it is to see the point of trying to make a difference when the chances are so remote. But the hybrid account shares with the expected consequences approach the advantage of giving a clear and compelling account of the relative strengths of

our reasons to contribute in various ways to various collective outcomes. While authors like Nefsky (2011; 2018; 2021) have challenged the cogency of this account, I draw upon defenses of the expected consequences approach (Barnett 2018, McMullen and Halteman 2019, Norcross 2020) to argue that it provides a more compelling account of the strengths of our reasons than Nefsky's development of the helping-based approach.

I argue, moreover, that the history and social science of attempts to create beneficial social change supports the conclusion that reasoning in accordance with the hybrid account provides substantial advantages in motivating and sustaining effective efforts at social change. The hybrid account foregrounds thinking about how one can most effectively help collective efforts to produce benefits, such as social and political movements, which historically shows to be essential for substantially beneficial social change (Teles and Schmitt 2011, Chater and Lowenstein 2022, Chibber 2022). Such maximally effective contributions include building solidarity, participatory strategizing, setting boundaries, and creating a healthy movement culture that can mitigate against burnout and biases. Studies of policy interventions, social movements, and unionization efforts find that these measures are critical for individuals to sustain their motivation and ability to take ongoing, strategically effective action, especially when they are acutely aware of the risks and uncertainties of doing so (Teles and Schmitt 2011, McAlevey (2016), Gorski et al. 2018, Centola 2021, Chibber 2022).

By contrast, the history of individual-frame policy interventions discussed by Chater and Lowenstein (2022) and the Effective Altruism movement demonstrate practical problems with using the expected consequences approach as a decision procedure. In seeking to optimize expected consequences, these interventions and this movement sought evidence of effectiveness in the form of easily measurable, tangible results from such things as randomized controlled trials. They shied away from efforts to create systemic changes, due to their perceived low tractability and difficulties in measuring impacts, but were in the end extremely naive and overconfident in the evidence-base for the efficacy of the apolitical interventions they favoured (Nathan 2016, Gabriel and McElwee 2019, Chater and Lowenstein 2022). Tendencies in the Effective Altruism movement have since endorsed efforts at system change (Berkey 2017, Kissel 2017, Matthews 2022), and Chater and Lowenstein can be understood as endorsing system change from an expected consequences perspective. But the bias in favour of tangible measurements of results and probabilities seems to be an inherent problem with using the expected consequences approach as a decision procedure (Ellsburg 2001), which using the hybrid approach instead can mitigate.

I conclude by showing how the hybrid approach can successfully address the problems Fanciullo (2019) raises for Nefsky's helping-based account about our reasons to contribute in cases involving mechanisms rather than other individuals. The hybrid approach can draw upon the resources it shares with

the expected consequences approach to argue that in these fanciful cases, one's act has a decent chance of helping to a degree, which constitutes a strong reason to contribute. This is counterintuitive, but only because the fanciful cases involve helping to bring about an outcome without a helping group, and our motivations to help are most powerful when acting in a group. By harnessing our powerful motivations to help in groups in real world cases of effective collective action, and insulating us from the cognitive biases of direct expected consequences reasoning, the hybrid approach is a more powerful and effective decision procedure.

Content moderation in Social Media Platforms

Paride Del Grosso

University of Antwerp

Abstract: Social media platforms (hereinafter SMPs), like Facebook and Twitter, have been increasingly using artificial intelligence (AI) to optimise content moderation. AI systems select contents published by SMPs' users and categorise these contents as ethically permissible or impermissible. Ethically impermissible contents are contents that are considered harmful (e.g. terrorist propagandistic posts, racist comments, etc.).

It has been questioned who has the right to decide which contents are ethically permissible or not (Sander 2020) and, consequently, whether the use of AI in content moderation should be regulated by governments or left to private self-regulation (Ferretti 2021). On the one hand, someone could claim that SMPs are private companies and, as such, their shareholders are the only who have the right to regulate SMPs according to their interests (self-regulation). On the other hand, SMPs have become so large-scale that they have assumed both a social and political nature (Bonini 2020, 265). Thus, although SMPs formally remain private agents that are independent from a government or a state (O'Neill 2001, 191), they have a huge influence on the public domain and have become, de facto, a digital agora where users can freely express their opinion. Hence, someone could also claim that it is wrong that the decision upon what it is ethically permissible or not is left in the hands of few people (i.e. the shareholders) and, thus, content moderation should be regulated by public and democratic entities or institutions, such as national governments.

In this paper, I will claim that content moderation in SMPs should be regulated by governments. I will support my claim by using an ethical argument based on the liberal-institutionalist approach, i.e. the ethical approach whose genesis can be found in Rawls' political theory (see Rawls and Kelly 2001). In sum, from the liberal-institutionalist approach I take the idea that democratic institutions (like governments) are the best actors to realise a just society, as they facilitate open decision making and have the democratic legitimacy to fairly implement the rules governing public life. The moral duty of single individuals and private actors (like companies) is to contribute to the justice of the society by respecting the laws and by developing improvements through public channels (Ibidem, 4).

By assuming this, I will conclude that, also in the case of content moderation in SMPs, regulations should be made by governments. The issue with content moderation is that, on one side, freedom of expression must be guaranteed, as it is a necessary feature of a just society (truth emerges from a 'free trade in the marketplace of ideas' (Brink 2001, 123)). However, on the other side, freedom of expression must also be limited in accordance to the harm

principle, i.e. freedom of expression should be banned when certain statements are harmful for specific categories of people (Ibidem, 121-122). In light of that, only governments, in virtue of their democratic legitimacy and transparent decisional processes, are entitled to decide what the right limit to freedom of expression, i.e. the limit of what it can be said or not..

I will challenge my position by considering two main objections. The first one comes from the individualistic ethical approach, according to which private agents have more legitimacy and capabilities than public institutions when it comes to make fair regulations and, hence, to decide where the limit to freedom of expression should be. The second one comes from the consequentialist ethical approach, according to which the maximisation of the aggregate utility (which represents the ethical purpose of a society) can be reached only if SMPs are self-regulated.

Lastly, I will discuss these two objections and I will conclude that, despite being compelling, they do not undermine the liberal-institutionalist argument that I present.

References:

Bonini, P. (2020). L'Autoregolamentazione dei Principlai Social Network. Rivista di Diritto Pubblico Italiano, Comparato, Europeo, 11: 264-281.

Brink, D.O. (2001). Millian Principles, Freedom of Expression and Hate Speech. Legal Theory, 7(2): 119-157.

Ferretti, T. (2021). An Institutionalist Approach to AI Ethics: Justifying the Priority of Government Regulation over Self-Regulation. Moral Philosophy and Politics.

Rawls, J., Kelly, E. (2001). Justice as Fairness: A Restatement. Cambridge, MA: The Harvard University Press.

Sander, B. (2020). Freedom of Expression in the Age of Online Platforms: The Promise and Pitfalls of a Human Rights-Based Approach to Content Moderation. Fordham International Law Journal, 43: 939-1006.

Overcoming the Particularity Problem

Avontay Williams

University of Alberta

Abstract: It is largely uncontroversial that there is some connection between voting and individual moral responsibility, but what is the connection? Julia Maskivker has recently attempted to justify a duty which requires individuals to vote well on the grounds of the good Samaritan principle, according to which if one can provide aid to others at very low cost to oneself, then one ought to do so. While this idea might seem to be neutral among theories of collective action, Maskivker thinks it is best supported by theories that reject appeal to expected consequences. This good Samaritan argument for voting well is challenged by Jason Brennan and Christopher Freiman who develop a concern for it which they call the "particularity problem," according to which the reasons usually given on behalf of a duty to vote fail to show there is a duty specifically to vote, but only at best show that voting is one of the many eligible ways to discharge some underlying duty, such as to exercise civic virtue, to avoid free-riding, to avoid complicity with injustice, or to maximize outcomes with high expected utility. For instance, instead of voting to positively change the outcome of an election, one might collect and donate to effective charities or pick up litter on election day. In this paper, I argue that the good Samaritan account of why we have a duty to vote well can overcome the particularity problem, but only by embracing the expected consequences approach (or something very similar to it). The expected consequence approach is the view that in collective action cases where our acts together are collectively beneficial, our individual acts do in fact have a chance (often small) of bringing about benefits (often large). As such, the moral importance of securing this chance of benefit typically outweighs the possible benefits to us of failing to perform these acts.

This paper's distinctive contribution to the existing literature is both theoretically and practically important. It is practically important, since by enabling us to appreciate the strength of our reasons to vote well as opposed to doing other things, I will show that voting well is morally obligatory. The other forms of easy aid that Brennan and Freiman suggest cannot be genuine substitutes for voting because these other actions cannot compete with voting given what's at stake in national elections. My argument is theoretically important, because in responding to the particularity problem, I contend that we really need a theory that addresses the strengths of reasons in various collective action cases in a plausible and compelling way. The expected consequences approach can do this, and it seems that rivals can't, at least without closely structurally imitating it. So, Maskivker is mistaken for thinking that we can defend duties to vote well without appealing to a

particular account of our reasons in collective action cases like the expected consequences approach.

I argue that, once we adopt an expected consequences framework, the plausibility of the idea that voting faces a particularity problem depends either on:

(1) The classical causal inefficacy problem, that the chances of one's vote for the superior candidate making a difference to the outcome of the election are too remote for the expected consequences of voting to be greater than the expected consequences of alternative courses of action, or

(2) A kind of implicit appeal to the idea that the chances of one's vote for the superior candidate making a difference to the outcome of the election are so remote that they are de minimis, or can be ignored for practical purposes – even if their mathematical expected consequences are greater than those of alternative courses of action.

In response to (1), I draw upon the work of authors like Zach Barnett (2020) and Andrew Gelman, Aaron Edlin, and Noah Kaplan (2008), who have successfully shown that a good analysis of the probabilities of votes making differences, together with a reasonable assessment of the stakes of national elections, shows the expected benefits of voting well to be substantial. I argue that because these expected benefits are almost always vastly greater than those of alternative courses of action that necessarily compete with voting, we almost always have a duty to vote well. I briefly mention and respond to an additional worry about the difficulty and costs of determining how to vote well that has been raised in a more recent paper by Brennan and Freiman (2022). While I leave elaboration of this response to another occasion, I briefly explain why I think (i) Brennan and Freiman selectively use unreasonably stringent standards of evidence, and why for most people most of the time (ii) the informational and deliberative costs of and cognitive biases against voting well are significant but not prohibitive or insurmountable.

In response to (2), I show that, when we consider the relevance of chances over a spectrum, from significant to very remote, a de minimis principle that departs from guidance by mathematically expected consequences is unreasonable.

AlphaGo, Intentionality, and the Prospect of Artificial Moral Patiency

Tuğba Yoldaş

University of Alberta

Abstract: In March 2016, Google's artificial intelligence (AI) computer program AlphaGo beat the world's 18-time Go champion Lee Se-dol, winning 4 of 5 games. Although the rules of Go are simple, it is a very complex game with 10360 possible configurations in a game of 150 moves (Granter et. al., 2016), "more than there are atoms in the universe" (Koch, 2016). The high complexity and intuitive nature of the game makes it inefficient to use brute force algorithms for a chance to compete against human players. Thus, AlphaGo is created with using machine learning algorithms, known as artificial neural networks, that simulate mammalian neural architecture to learn from millions of games played by expert human Go players by studying their game positions. Further, to become a better player than an expert human, AlphaGo played millions of games against itself thereby learning and improving through "reinforcement learning" (Silver et. al., 2016; Granter et. al., 2016): the program simply taught herself which moves brought up better outcomes, i.e., wins, or moves that lead to wins. With AlphaGo, some thought that humans created truly thinking machines who are much smarter or more "intelligent" than us, at least in areas like the game of Go. Moreover, Google's DeepMind extended their project to creating smarter machines, such as AlphaGo Zero (Silver et. al., 2018). In simplest terms, AlphaGo Zero doesn't learn and improve its game by studying expert human players' moves, but rather she teaches herself by self-play, with no human supervision, and using only raw board history as input (Silver et. al., 2018). There is something eery about all these achievements. Nobody, even the expert human Go-players, knows why the program plays so well, e.g., see the discussions around the unimaginable "move 37" (see., e.g., Halina, 2021) and the surge in XAI (The Explainable AI): "AlphaGo is the creation of humans, but the way it plays is not" (Roeder, 2016).

In this paper, I will first ask the question, "What is the best strategy to make sense of the behavior of any system?", and I will start with the most familiar case: humans. When we act with other people, we often rely on our understanding of intentional states like beliefs, desires, intentions, motivations, hopes, and concerns. We take what Dennett (1971; 1981; 1987; 1989) calls the "intentional stance" toward other people in order to predict, explain and plan actions in navigating through our social relationships. According to Dennett (1997), there are three main stances you may adopt to explain and predict the future behavior of systems: physical, design and intentional stances. Consider the example of seeing me turning off the lights. Adopting the physical stance, you may try to make sense of and predict what

I will do next and why in the first place I turned off the lights by appealing to certain physical laws. For example, you may say, “She turned off the lights because the brain area X went active and there was electrical activity between such-and-such types of neurons that sent a message to the primary motor cortex”. This stance surely doesn’t explain my action, nor does it give you predictive power to explain my next action. It is simply impractical and impossible to explain my behavior from the physical stance. Similarly with the design stance: appealing to how humans work as they are meant to by design, e.g., natural selection mechanisms, doesn’t make sense of our actions. For one (among many), evolutionary selection mechanisms are argued to be not “fine-grained” enough to distinguish between one mental content over another in most cases (Fodor, 1990), leaving biological functions causally indeterminate (cf.

Millikan’s (1989) biosemantics). Dennett argued that the only practical and possible way to predict and make sense of the behavior intentional systems, such as us, is to adopt the intentional stance by assuming that rational agents will behave in accordance with their beliefs, desires, intentions in order to achieve certain goals. You may make sense of my behavior by saying, for instance, “She turned off the lights (with the intention to sleep) because she wants to sleep and believes that turning off the lights will fulfill her desire to sleep”.

Next, I will discuss what it would take for an AI system to have the intentional states akin to the kind of intentional states that humans and non-human animals have. In this paper, I will evaluate the intentionality of AlphaGo and its successors with the aim to understand what kind of features would confer an AI system like AlphaGo the status of a basic moral patient (see, e.g., Basl, 2013 for a discussion on moral patiency). In Nye and Yoldas (2021), we argued that for an AI system to be a basic moral patient to whom we can owe duties of maleficence not to harm her and duties of beneficence to benefit her, the system must be a mental patient who is capable of mental states like experiences, beliefs, desires and motivations. We also argued that mental patients are true intentional systems whose behavior can be best explained by appealing to its representations and goals that can flexibly interact with a wide variety of the system’s other representations and goals in the way that the philosophical theory of success semantics described (Whyte, 1990; 1991): the theory roughly explains the content of beliefs or their truth conditions in terms of the fulfilment conditions or the content of desires, and proposes that some of our desires have contents that are not explained by reference to the content of beliefs. There is a sense in which we can attribute to AlphaGo the goal of winning a Go game, variety of intermediate goals along with a variety of representations that will achieve those goals and explain her behavior from the intentional stance. However, I will argue that AlphaGo exhibits “real patterns” of behavior that can be interpreted as having truly determinate representational contents only relevant to the playing of Go while it lacks domain general intelligence. We interpret AlphaGo’s states as being about Go,

winning, capturing, pieces because those are the purposes with which AlphaGo is designed, or which we have interacting with the system. AlphaGo Zero has more systematically flexible representations and goals, but still lacks domain-generality. I conclude the paper by arguing that this is a good thing because we have strong moral reasons not to create artificial moral patients at least for now, and I discuss some relevant ethical implications of creating artificial moral patiency.

Epistemic Hierarchy

Idowu Odeyemi

University of Colorado Boulder

Abstract: Epistemic hierarchy, as I conceptualize it in this paper, refers to a particular social phenomenon, whereby an individual or group of people pervasively engage in practices of epistemic vice, hence tentatively building structures of epistemic dominance in order to discredit the epistemic capabilities and/or epistemic utility of the assumed lower epistemic group or individual. Practices of epistemic vice, in social relations, can operate actively or passively. The concept of epistemic hierarchy, when it manifests in social relations, assumes an epistemic order of knowing and gives an inflationary account that is coextensive with the class of a priori knowable claims. That is, in its social manifestation, epistemic hierarchy latently stipulates "intelligence" as objective, in a form of a Platonic ideal, and offers an account that is incompatible with pragmatic reasons.

I identify two forms in which epistemic hierarchy takes place: the first is epistemic arrogance. I quantify it within individual social relations owing to prejudice, most times, taking place as a sort of what Mathew Cull (2019) calls dismissive incomprehensibility—an attitude that comes up when a hearer is unwilling to fundamentally engage with epistemic virtue and understand that her lack of comprehension of the speaker's testimony is not due to the epistemic credibility of the speaker but a purported lack of access to the speaker's cultural, social, and personal reality. When dismissive incomprehensibility happens in social relations between individuals, the hearer fails to assign credibility to the coherence of the speaker's epistemologies. Dismissive incomprehensibility is not the only form of epistemic arrogance. Another form of epistemic arrogance comes up when the hearer fails to assign epistemic credibility to the speaker's testimonies because of the social-cultural history of the speaker's epistemic situation. This second form of epistemic arrogance has been discussed by Miranda Fricker (2007) as testimonial injustice. However, I show how this second form might happen consciously or subconsciously in individual social relations.

The second form of epistemic hierarchy is epistemic dominance. I identify it within social relations owing to the power and the quantity of upholding epistemic arrogance within a group of people. That is the wrong-making process of epistemic dominance occurs when the ideologies birthed by epistemic arrogance are coded into the norms of a society. The credence of badness is what births epistemic dominance. Even though epistemic dominance, an oppressive event, usually results in what Abraham Tobi (2021) calls appreciative silencing, a phenomenon where the accepted hegemonic intuitions of the oppressed are formed by the oppressors' ideologies over time

when internalized, I argue that the epistemic silencing elicited by epistemic dominance, via the assumption of an epistemic hierarchy, is purposefully harmful. Epistemic hierarchy is harmful, especially to epistemic diversity, an agenda that should be used to drive-out epistemicide, the destruction of coherent knowledge within a group's social life. For example, the most devastating effect of epistemic colonialism is that it sets up a single perspective, in which the epistemic credibility of the oppressed is dismissed due to the establishment that the oppressors, often Western, epistemic practices are more reliable. Also, the prejudicial claim that "men's intuitions are more reliable than those of women" over time sets the foundation for gender oppression.

The distinction between the two forms of epistemic hierarchy stems from individual relations: Epistemic arrogance grows into a more harmful stage of epistemic dominance when the ideologies epistemic agents prejudicially converge on are coded into the norms of a society. Credence of badness, then, distinguishes epistemic arrogance from epistemic dominance. For instance, normalizing individual prejudices against the epistemic credibility of Africans are what leads to colonialism. These two—epistemic arrogance and epistemic dominance—are pernicious epistemic forms that lead to epistemicide—the event that leads to different forms of social and political problems such as racism, sexism, and elite capture among others. Epistemic arrogance is about prejudice. Epistemic dominance is about power. Distinguishing between epistemic arrogance and epistemic dominance will make explicit the latent structures of the procedure of epistemic to social oppression that has coordinated our social existence over time.

However, at this point, it is important to make clear that epistemic hierarchy is not necessarily a bad or negative social phenomenon. There are cases in which practices of deference in an epistemic hierarchy are appropriate. That is in some situations it is appropriate for an epistemic agent to practice epistemic deference by deferring to an epistemic authority that warrants epistemic trust. Some of these cases occur on a professional level. But establishing epistemic diversity is necessary to make adequate the credibility of epistemic authority in a globalized world.

This paper will then proceed as follows. In section two, I conceptualize epistemic hierarchy. Here, I first define the concept of epistemic hierarchy. Then I argue that there are circumstances in which epistemic hierarchy is appropriate. But my focus is on the inappropriate role epistemic hierarchy plays in the domain of the social-epistemic phenomena. I identify that this negative form of epistemic wrong comes into play in two forms: epistemic arrogance and epistemic dominance. My focus is on these two negatives. In section three, I offer an account of the appropriateness of epistemic hierarchy in social structure. I argue that understanding the structures of appropriate social elicitation of epistemic hierarchy is required to illuminate the two negatives of epistemic hierarchy. In section four, I give an explicit account of

how epistemic hierarchy manifests inappropriately. First, I define the problem by giving an analysis of its latent structure over time in human history. I then account for the two inappropriate forms of epistemic hierarchy: epistemic arrogance—given the prejudicial label—and epistemic dominance—given the power the convergence of epistemic arrogance establishes. Section five is entitled Tracking Epistemic Hierarchy, Tracking Practices of Epistemicide. Here, I give an analysis of how an assumption of epistemic hierarchy has over time established social wrongs such as racism, colonialism, sexism, and what Cynthia Townley (2003) calls harm to one's "epistemic agency." In this section, I also account for how epistemic differences make it wrong that epistemic hierarchy persists. In section six, I examine if an African model of epistemic democracy can offer a solution to the problem at hand: can the epistemic discursiveness' that is associated with African epistemic democracy clear our sight of the stimulating pursuit of a total answer to the question "what is the truth?" In section seven, I reply to three objections: expertise defense, accessibility defense, and objectivity defense. In section eight, I conclude by offering insights into two bewildering facts that inspired this paper. First, the hardly unrealized structures of the negative epistemic hierarchy are indebted to the objective standard set for knowledge. And secondly, these negatives form the edifice of all forms of social and epistemic wrongs such as racism, sexism, classism, epistemic subjugation, and epistemic silencing, among many others.

References

Bokros, Sofia Ellinor (2020). A deference model of epistemic authority. *Synthese* 198 (12):12041-12069.

Fricker, Miranda (2007). *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford University Press.

Cull, Matthew J. (2019). Dismissive Incomprehension: A Use of Purported Ignorance to Undermine Others. *Social Epistemology* 33 (3):262-271.

Tobi, Abraham (forthcoming). Appreciative Silencing in Communicative Exchange. *Episteme*:1-15.

Townley, Cynthia (2003). Trust and the Curse of Cassandra. *Philosophy in the Contemporary World* 10 (2):105-111.