social sciences

MDPI

Review

An Exploration of Ethical Decision Making with Intelligence Augmentation

Niyi Ogunbiyi *, Artie Basukoski and Thierry Chaussalet

School of Computer Science and Engineering, University of Westminster, London W1W 6UW, UK; A.Basukoski@westminster.ac.uk (A.B.); chausst@westminster.ac.uk (T.C.)
* Correspondence: oluniyi.ogunbiyi@my.westminster.ac.uk

Abstract: In recent years, the use of Artificial Intelligence agents to augment and enhance the operational decision making of human agents has increased. This has delivered real benefits in terms of improved service quality, delivery of more personalised services, reduction in processing time, and more efficient allocation of resources, amongst others. However, it has also raised issues which have real-world ethical implications such as recommending different credit outcomes for individuals who have an identical financial profile but different characteristics (e.g., gender, race). The popular press has highlighted several high-profile cases of algorithmic discrimination and the issue has gained traction. While both the fields of ethical decision making and Explainable AI (XAI) have been extensively researched, as yet we are not aware of any studies which have examined the process of ethical decision making with Intelligence augmentation (IA). We aim to address that gap with this study. We amalgamate the literature in both fields of research and propose, but not attempt to validate empirically, propositions and belief statements based on the synthesis of the existing literature, observation, logic, and empirical analogy. We aim to test these propositions in future studies.

Keywords: ethical decision making; explainable AI; Intelligence augmentation; Values in Design

Citation: Ogunbiyi, Niyi, Artie Basukoski, and Thierry Chaussalet. 2021. An Exploration of Ethical Decision Making with Intelligence Augmentation. *Social Sciences* 10: 57. https://doi.org/10.3390/socsci10020057

Academic Editor: Feras A. Batarseh
Received: 8 January 2021
Accepted: 3 February 2021
Published: 8 February 2021

1. Introduction

The use of Artificial Intelligence (AI) agents has gained widespread attention in the last few years (Science and Technology Committee 2018). As used in this paper, AI refers to "a set of statistical tools and algorithms that combine to form, in part intelligent software enabling computers to simulate elements of human behaviour such as learning, reasoning and classification" (Science and Technology Committee 2018). One of the prominent uses of AI is to assist human stakeholders in decision making (Abdul et al. 2018). This has been described as "Intelligence augmentation", as AI models are used to "help improve the efficiency of human intelligence" (Hassani et al. 2020). As highlighted by the Academy of Medical Science (2017), Intelligence augmentation has been used in healthcare to enable "clinicians work more efficiently and better handle complex information". It has also been utilised in the criminal justice system to detect crime hotspots and decide whether a suspect could be eligible for deferred prosecution (Oxford Internet Institute 2017), and by financial services providers to determine the outcome of a credit application (Financial Service Consumer Panel 2017), amongst others. Intelligence augmentation has resulted in significant benefits including improved quality, more personalised service, reduced processing time, and more efficient allocation of resources.

However, several issues have arisen that have raised a cause for concern. For example, several high-profile instances have been highlighted where similar individuals with identical financial data, but different gender have had different outcomes to credit applications (Peachey 2019). Allegations that AI algorithms used in the criminal justice system discriminated against defendants based on race have also been raised (Maybin 2016). This

algorithmic bias is attributed to unrepresentative or insufficient training data, sophisticated pattern learning which can discover proxies for protected characteristics (e.g., gender, race, sexual orientation, and religious beliefs)—even when these are explicitly removed from the data, amongst others (see Bell 2016; Murgia 2019). The issue has gained such attention that the UK Parliament Select Committee on Science and Technology commissioned an enquiry to investigate accountability and transparency in algorithmic decision making (see Science and Technology Committee 2018). The IEEE Standards Association also introduced a global initiative for ethical considerations in the design of autonomous systems (see IEEE 2016). The Association for the Advancement of Artificial Intelligence (AAAI) in its code of conduct acknowledged that "the use of information and technology may cause new or enhance existing inequalities" and urges "AI professionals [. . .] to avoid creating systems or technologies that disenfranchise or oppress people" (see Association for the Advancement of Artificial Intelligence 2019).

In terms of positioning this study, we briefly discuss related studies. The study by Paradice and Dejoie (1991) established that "the presence of a computer-based information system may influence ethical decision making". However, we presume that given that this study predates the recent exponential growth in the capability and ubiquity of AI tools, it does not address the peculiar challenges of AI tools in ethical decision making. Johnson (2015) advances the topic to include artificial agents, highlighting the "push in the direction of programming artificial agents to be more ethical". Martin (2019) extends the discussion further, positing that algorithms are "not neutral but value-laden in that they [. . .] reinforce or undercut ethical principles" and highlights that "algorithms are [. . .] an important part of a larger decision and influence the delegation of roles within an ethical decision". Martin et al. (2019) argue that "ethical biases in technology might take the form of [. . .] biases or values accidentally or purposely built into a product's design assumptions".

Objectives

This paper aims to contribute to the literature base by synthesising the fields of ethical decision making and Explainable AI (XAI). It puts forward, but does not attempt to validate empirically, propositions and belief statements that can be subsequently tested in future studies. These propositions are based on conclusions derived from the existing literature, observation, logic, and empirical analogy. The scope of the study is Intelligence augmentation, where an AI model makes a recommendation to a user (who makes the final decision) as opposed to automation, where autonomous machines make decisions previously entrusted to humans.

A better understanding of how users navigate these ethical issues is of interest in evaluating decisions made by human agents using AI models regardless of the degree of transparency of the model. Martin (2019) argues that "responsibility for [. . .] design decisions [which allow users to take responsibility for algorithmic decisions] is on knowledgeable and uniquely positioned developers". By shedding light on how human agents make decisions with AI models, it would also assist developers with the design of explainable AI (XAI) systems that would assist human agents in identifying ethical issues and dealing with them appropriately. This will serve to improve Intelligence augmentation, which will only increase as more AI tools are deployed in "the wild" (Ribeiro et al. 2016).

Enhanced understanding of the human decision-making process with AI would also benefit policy makers and regulators who are increasingly focused on protecting "the public from discrimination by algorithms that influence decision-making on everything from employment to housing" (Murgia 2019).

The issue is relevant and salient as it assists with answering questions about accountability, i.e., who is responsible when a human agent accepts an unethical recommendation proposed by an AI model: Is it the human decision-maker or the AI agent? The UK Parliamentary Select Committee report recommends exploring "the scope for individuals [...] where appropriate, to seek redress for the impacts of such decisions". Some experts are

"wary of placing full responsibility on the user of an algorithm" (Klimov 2017). That would suggest that a degree of responsibility (however small) rests with the user. Other experts suggest that "we may want to assign strict liability [to the user of the algorithm] in certain settings" (see Weller 2017).

Two factors compound this issue further:

(1) Human users tend to assign traits typically associated with other humans (e.g., intentionality, beliefs, desires) to AI tools (De Graaf and Malle 2017).
(2) The acknowledgement that these models can process vast amounts of data effectively and discover interactions in the data far beyond a typical human's comprehension (Amoore 2017).

The combination of these factors increases the likelihood that an unethical recommendation by an AI model will be accepted as it is regarded as a trusted expert.

The remainder of the paper is structured as follows: Section 2 defines vital terms built on throughout the paper. Section 3 discusses the basis for the findings and propositions from the literature synthesis, while the final section summarises recommendations and proposes further research areas for extending these recommendations.

2. Definitions

Several key terms to be developed further and built on throughout this paper are defined in this section.

Moral agent:

A person who makes a moral decision regardless of how the issue is constructed (Sonenshein 2007). In the context of this study, the moral agent is the stakeholder who decides with the aid of an AI model. For example, the Human Resources (HR) officer who determines that a job application should not proceed based on the recommendation of a model. The moral agent is also referred to in this paper as "the user" of the AI agent.

Ethical decision:

Several studies have highlighted the lack of a widely accepted definition of ethical behaviour (see Cavanagh et al. 1981; Beauchamp et al. 2004). Rather than base our definition of an ethical decision on consensus (e.g., see Jones 1991; Treviño et al. 2006), we adopt definitions based on a priori principles, e.g., Kant's (1785/1964) respect principle (see Tenbrunsel and Smith-Crowe 2008). For example, it is unethical to disrespectfully discriminate against a person based on their ethnicity or gender, while the converse is also true. Smith-Crowe (2004) and Bowie (2017) provide further examples of how Kant's principle is applied in business. For the purpose of this paper, morality and ethicality are used interchangeably. In other words, a morally "correct" decision is considered an ethical decision and vice versa (see Jones 1991).

Tenbrunsel and Smith-Crowe (2008) argue that unethical decisions could be made intentionally or unintentionally (intended and unintended unethicality). We posit that the use of AI models has the potential to significantly increase instances of unintended unethicality where a human agent accepts the recommendation of a model without realising it may be flawed.

Explainability:

The ability of an AI model to summarise the reason for its behaviour or produce insights about the causes of its decisions. Explainable models are also described as "transparent" models. Closely associated with explainability is the quality of explanation, i.e., is the explanation "good" enough? Gilpin et al. (2018) posit that the quality of the explanation can be evaluated by its degree of interpretability and completeness.

Explainer:

An agent who supplies an explanation for the recommendation made by itself or another AI model. Examples of explainers from the literature base include LIME (Ribeiro et al. (2016), OC-DTD (Kauffmann et al. 2020), and PJ-X (see Park et al. 2018). Xu et al. (2019), Adadi and Berrada (2018), and Li et al. (2020) are among many studies which present a detailed survey of explainers.

Explainee:

A person to whom an explanation is supplied, often in response to a request for an explanation. In this context, the explainee is usually the moral agent who makes the final decision based on the recommendation provided by the AI model.

Interpretability:

The ability to describe what the AI model did (or did not do) in a manner that is understandable to users. As users vary in their level of skills and expertise, interpretability requires the ability to describe, in a flexible and versatile manner, tailored to the user's particular mental model. Ribeiro et al. (2016) make a connection between a user's "trust" in the system and the likelihood of accepting the prediction of the model. They make a distinction between trusting a prediction and the model as a whole. Trust at both levels is predicated on how much the user understands the model's behaviour.

Completeness:

The ability to describe the operation of an AI model accurately. An explanation is more complete if it allows the behaviour of the model to be anticipated in more situations (Gilpin et al. 2018).

3. Literature Synthesis

We commence by examining rationalist (or reason-based) models of ethical decision. Numerous models have been proposed that view ethical decision making as a rational process (Ferrell and Gresham 1985; Rest 1986; Trevino 1986; Hunt and Vitell 1986). Perhaps the best-known of these models is that proposed by Rest (1986). This model argues that ethical decision making progresses through a four-stage process from recognising a moral issue ending with engaging in moral behaviour. Jones (1991) extended this model further to develop an issue-contingent model which argued that the moral intensity of an issue influenced ethical decision making and behaviour. However, Sonenshein (2007) highlights four key limitations of rationalist models as follows: (i) they fail to adequately address the presence of equivocality (i.e., the existence of multiple interpretations) and uncertainty (i.e., lack of complete information) that are present in many real-world scenarios; (ii) they presume that ethical behaviour is preceded by deliberate reasoning; (iii) they fail to fully emphasize the construction of ethical issues; and (iv) they assume a strong causal link between moral reasoning and judgement. We recognise that we must build on a model that addresses these limitations. For example, due to varying degrees of transparency, equivocality and uncertainty are ubiquitous in decision making with AI tools. Hence, we adopt the Sensemaking Intuition Model (see Sonenshein 2007), which addresses these limitations, as the foundation on which we build our synthesis.

However, we will first consider how a human user interacts with an explainable AI agent to obtain explanations and subsequently arrive at a decision (ethical or otherwise—see Figure 1). The process commences when the user detects that a recommendation it has received from the AI agent is abnormal. The user subsequently evaluates the explanation provided by the explainer and selects a subset of it. Depending on how comprehensible or plausible the explanations are, the user may request clarification, which is, in turn, evaluated. The user may conclude the explanation/clarification is plausible and accept the recommendation or conversely may conclude that the recommendation provides evidence of algorithmic bias and reject the recommendation.

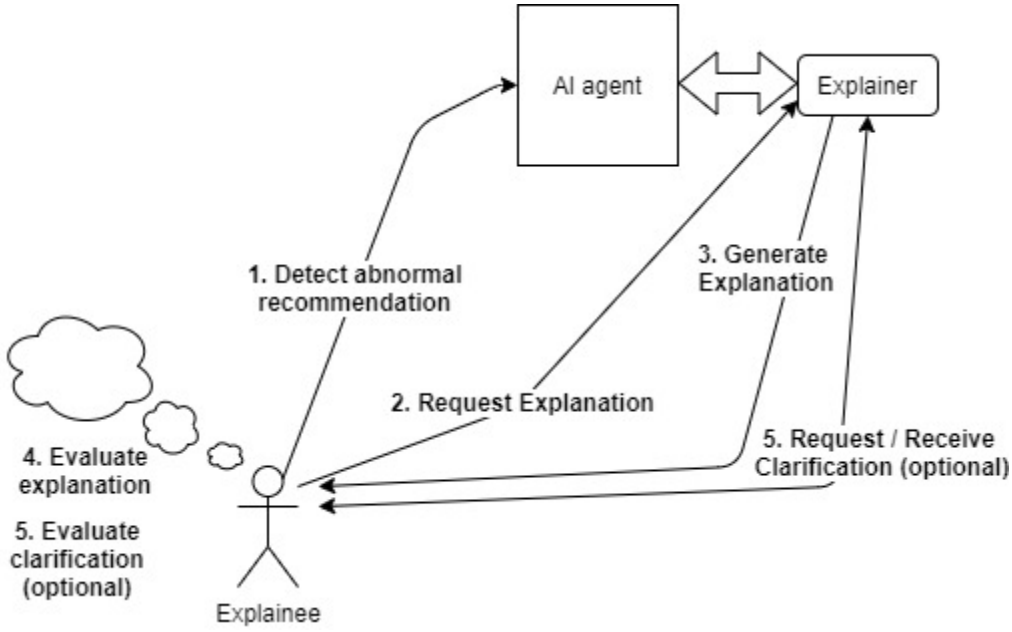

Figure 1. The AI explanation cycle.

Figure 2 maps the AI explanation cycle to the various stages of the ethical decision-making model. Though this is not a precise mapping—for example, the user may request clarification after making an intuitive judgement—the mapping is useful for designing interventions that will increase the likelihood that a user will detect algorithmic discrimination and behave ethically.

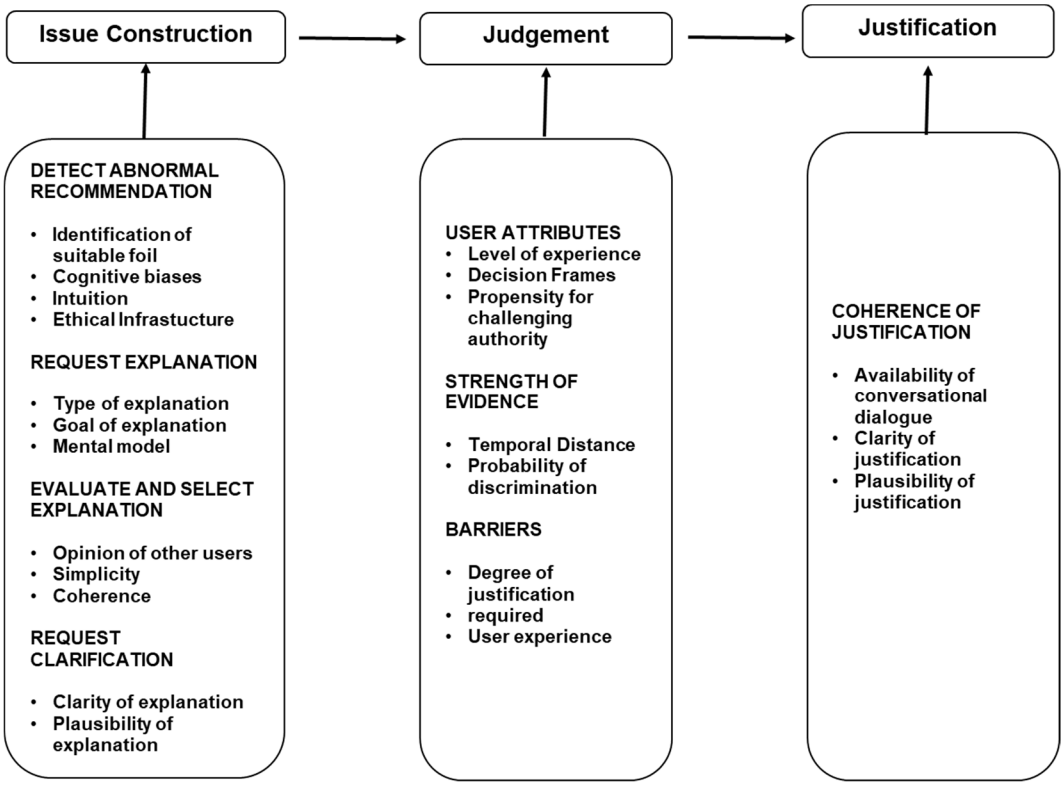

Figure 2. Synthesised model of ethical decision making with Intelligence Augmentation.

3.1. Issue Construction

3.1.1. Expectations

Sonenshein (2007) posits that the ethical decision-making process commences with issue construction where "individuals create their own meaning from a set of stimuli in the environment".

We argue that in the context of AI tools augmenting human judgement, the issue construction process is influenced by the degree of transparency of the AI model. Sonenshein argues that "individuals' expectation affect how they construct meaning". In the case of an opaque (or black box) model, whether the user recognises the recommendation as "abnormal" will depend on the identification of a suitable "foil". As established by several studies, people tend to request clarification about observations that they consider unusual or abnormal from their current perspective (see Hilton and Slugoski 1986; Hilton 1996). Van Bouwel and Weber (2002) argue that establishing abnormality is often done using a contrastive case (also referred to as a foil). Of particular interest in this regard is what they label the O-contract of the form: why does object a have Property P, while object b has property Q? To be more precise, we consider the case where object a has Property P = X, while an identical object b has Property P = Y. Consider two individuals, a and b, identical in all respects except for gender, who both submit a loan application around the same time. However, the model recommends that one application be approved, while the other is denied. If the user is aware of the recommended outcome in both cases, one of the cases will serve as a foil (or counterfactual). The user will utilise abductive reasoning to attempt to determine the cause of the observed recommendation (see Peirce 1997). To accomplish this, the user will generate several hypotheses as to the likely causes of the recommendation (one of which is likely to be that there is algorithmic discrimination at play), assessing the plausibility of these hypotheses and selecting the "best" hypothesis. Harman (1965) describes this process as "inference to the best explanation". If algorithmic discrimination is thought to be the best hypothesis (regardless of whether or not it is the real cause), the user will construct the issue as ethical. However, there may not be a foil readily available, or the user may not be aware of it, in which case, the user may not construct the issue as an ethical one.

In the case of an AI model with any degree of transparency, the trigger for detecting abnormality typically starts with a request for an explanation. Though Miller (2019) posits that curiosity is the primary reason an explanation is requested, we argue that in this context, an explanation is more likely to be requested for regulatory or customer relations management reasons, i.e., to justify the decision made to a regulator or customer, respectively. However, the issue construction process is dependent on how the system presents the reasons for the recommendation as well on how the user selects and evaluates the explanation. Though most transparent models present their explanation as causal chains or probabilistic models, Miller (2019) argues that "whilst a person could use a causal chain to obtain their own explanation [...] this does not constitute giving an explanation". In terms of explanation evaluation, he argues that "whilst likely causes are good causes, they do not correlate with explanations people find useful". He posits from his review of the literature that there are three criteria people find useful in evaluating explanations: simplicity, generality, and coherence. To illustrate this, consider the case of a user requesting an explanation for a recommendation from an explainer such as LIME (Ribeiro et al. 2016). The user is presented with a list of features that contributed to the recommendation in the order of magnitude of their contribution. The user subsequently assesses whether or not the features that drove the recommendation are plausible based on their subject matter expertise. If some unexpected features are driving the recommendation, the user may view this as an abnormal recommendation. The user may search for a foil (i.e., a similar case), examine whether a similar recommendation was given and whether similar unexpected features drove this. If the identical cases have different outcomes, the user may be alerted to the existence of an ethical issue.

Given the contrast between opaque and transparent models as described above, we argue that the more transparent a model is, the more likely it is that the agent will construct the issue as an ethical one. For example, if the user can understand which features made the most significant contribution to a prediction, they are more likely to detect if algorithmic discrimination exists (interpretability). By the same token, because a more complete explanation is likely to shed light on of the system's behaviour, it is likely to make the user recognise the existence of ethical issues than a less complete system.

Proposition 1. *The more explainable an AI model is, the more likely it is that the human agent will construct the issue as ethical as compared to less explainable models.*

3.1.2. Explanation Goal

We argue that the type and goal of the explanation requested also influences the issue construction process. Initially, the user may request an explanation for the abnormal recommendation vis-à-vis the foil. However, where the issue is of high moral intensity (e.g., see Maybin 2016) or there have been repeated instances of abnormal recommendations, the user starts to question the credibility of the entire system. Rather than request an explanation for a specific recommendation, they start to ask for explanations about the model itself and its learning configuration—a "global perspective" which explains the model (see Ribeiro et al. 2016). This requires a "model of self" which approximates the original model and exists primarily for an explanation (see Miller 2019). That paper highlights an example of such an explanatory modification of self from a study by Hayes and Shah (2017).

Numerous studies have demonstrated that the user evaluates and selects a subset of explanations provided by the explainer as relevant based on factors such as abnormality, the contrast between the fact (i.e., observed recommendation) and the foil, and robustness, amongst others (see Miller 2019). Also related to issue construction, the study by Kulesza et al. (2013) explored the link between the soundness (or correctness) and completeness of the explanation. They recommended that while completeness was more critical than soundness, it was important not to overwhelm the user. Miller (2019) also argues that when the entire causal chain is presented to the user, there is a risk that the less relevant parts of the chain will dilute the crucial parts that are important to explain the recommendation. This recommendation runs contrary to the intuitive view that more information is better than less.

Proposition 2. *The more interpretability (as opposed to complete) the supplied explanations have, the more likely it is that the human agent will correctly construct the issue as a moral one.*

3.1.3. Motivational Drive

Tenbrunsel and Smith-Crowe (2008) argue that "biases, intuition and emotion must be considered" in the ethical decision-making process. This aligns with the position put forward by Sonenshein (2007) that "individuals see what they expect to see, but [. . .] also see what they want to see". We consider the implications of these biases on ethical decision-making using AI tools.

Messick and Bazerman (1996) postulate that internal theories influence the way we make decisions. Strudler and Warren (2001) provide an example of one such bias (authority heuristics) which describes the trust we place in the expertise of authority figures that may be misplaced. As we argued earlier, humans tend to view AI models as authority figures. However, as highlighted earlier, the user is likely to bring biases into the evaluation of explanations provided based on their perceived intention of the explainer (see Dodd and Bradshaw 1980).

Abnormality is a critical factor in ethical decision making as it triggers the request for an explanation regarding the basis for the model's recommendation. (Miller 2019). However, what is viewed as abnormal is subject to cognitive biases held by the user. For

example, Gilbert and Malone (1995) highlight correspondence bias due to which people tend to explain other people's behaviour based on traits. In other words, a user may not view a model's recommendation as abnormal due to discriminatory tendencies they may harbour or may give higher weight to unimportant causal features that support biases. It is also possible that the user may select a conjunction of facts in the causal chain and assign them higher weighting that they deserve because it aligns with their preconceptions (see Tversky and Kahneman 1983).

Proposition 3. *The more aligned a human agent's biases are with the AI model's recommendations, the less likely they are to construct the issue as a moral one.*

3.1.4. Social Anchors

We posit that the existing "ethical infrastructure" in the organisation is another important factor that impacts the issue construction process when making decisions with AI tools. Ethical infrastructure refers to "organisational climate, informal systems and formal systems relevant to ethics" (Tenbrunsel et al. 2003). Where the ethical infrastructure supports constructing moral issues regarding the existence of algorithmic discrimination, the user is likely to do so; otherwise, they will not.

Sonenshein (2007) argues that "employee's goals [. . .] will affect how they construct an issue". They may view unethical behaviour as consistent with the "rules of business" if it enables them to achieve their goals. This is consistent with results from Schweitzer et al. (2004), which concludes that goal setting is negatively associated with ethical behaviour. This conclusion aligns with the findings by Hegarty and Sims (1978) and Tenbrunsel (1998), which discovered a positive correlation between incentives and unethical behaviour. In terms of decision making with AI tools, this suggests that if an organisation sets goals that encourage specific outcomes based on AI tools without taking appropriate action to manage undesirable side effects, users are less likely to construct ethical issues appropriately.

Proposition 4. *Users working with AI models in organisations with more supportive ethical infrastructures are more likely to challenge the model's recommendation compared to users in organisations with less supportive ethical infrastructures.*

3.1.5. Representation

Sonenshein (2007) posits that the user's representation—their "mental model [. . .] of how others see a situation"—is also an important moderator of issue construction. We argue that a significant "other" in decision making with AI tools is the explainer which can shed light on the factors that drove the recommendation made by the AI agent. Sonenshein quotes a study by Weick (1993), which found that people engage in representation through communication with others. This highlights the need, not only for the user to be able to request an explanation, but also about the nature of the explanation provided. For example, though the explanation provided could include details such as the training data, optimisation cost function, hyperparameters, etc., these are not likely to be useful to the typical user. It is worth noting that the user is likely to be distrustful of the explanation offered. This conclusion is based on research which indicates that individuals tend to prejudge the intention of the explainer and filter out information that supports the prejudgement (see Dodd and Bradshaw 1980). As a result, the user is likely to request clarification and additional explanation of any explanations provided. This supports the requirement for an explanation as dialogue, which facilitates challenges from the explainee ("I do not accept your explanation or parts of it").

The other form of representation that is relevant is the opinion of other users of the AI tool. Sonenshein (2007) refers to these as "social anchors, . . . interlocutors who help an actor test his or her interpretation of social stimuli". This suggests that the way a user constructs the ethical issue in isolation is likely to be different from the manner it will be constructed if done collaboratively with other users. In the latter scenario, additional users

are likely to be able to provide additional examples of foils which will broaden the initial user's frame of reference and enable them to construct the issue in a broader manner.

Proposition 5. *The more collaborative the issue construction process is, the more likely it is that the user will correctly construct an issue as ethical as compared to issue construction done in isolation.*

3.2. Judgement

3.2.1. User Attributes

Sonenshein (2007) posits that the intuitive judgement stage directly follows the end of the issue construction stage with the user reaching a plausible interpretation. At this stage, the actor responds to the issue (as constructed in the initial stage) using intuition—an "automatic, affective reaction". With regard to ethical decision making with AI tools, the agent makes an intuitive judgement which rules the recommendation as discriminatory or non-discriminatory.

Sonenshein (2007) further argues that an individual's level of experience is a key factor that influences their judgement, stating that "as individuals develop experience, they can internalise that experience into intuitions". Regarding ethical decision making with AI tools, this implies that a less experienced user is likely to challenge the recommendation of the AI tool. They are likely to have less experience of abnormal recommendations and as such, are more likely to view the AI tool as an expert, the converse of which is true of more experienced users.

Also relevant at this stage is the work of Tenbrunsel and Smith-Crowe (2008). The authors introduce the concept of decision frames that illuminate the perspective of the decision-maker and are moderated by previous experience. For example, suppose the decision-maker primarily adopts a business or legal frame. In that case, they are less likely to judge a model's recommendation as discriminatory (even if there exists evidence to the contrary). Though Sonenshein suggests that individuals infrequently alter their initial judgement after it has been made, we argue that exposing the basis on which an AI model made a prediction can shift the user's decision frame such that what may have been perceived at the outset as a business/legal decision is transformed into an ethical one. We believe this is especially the case for more morally intense issues, e.g., if the AI model is being used to determine the risk of reoffending for an offender (Maybin 2016).

We argue that the degree of interpretability also influences the user's judgement. For example, given an opaque model and the absence of a suitable foil, the user is unlikely to judge the model's recommendation as discriminatory. This is backed up by findings which relate the degree of perceived control over an event with attribution of responsibility (Fiske and Taylor 1991). In other words, a user making a decision based on a prediction from a black box model is likely to attribute the decision to the model ("Computer says 'No'") as opposed to a prediction based on an interpretable model where the user is more likely to perceive that they have more control.

Kelman and Hamilton (1989) argue that an individual's propensity for challenging authority (i.e., the model's recommendation) depends on which is the more powerful of two opposing forces in tension—binding and opposing forces. Binding forces strengthen the authority of existing structures while opposing forces intensify resistance to authority. As stated earlier, human users tend to view the model as the "authority" due to their data computation ability. We argue that interpretability is likely to heighten the opposing force and make the user more likely to challenge the recommendation where it is abnormal. In addition, it will also make it easier for the user to justify their rationale for disregarding the model's recommendation.

Proposition 6. *The more experienced a human agent is, the more likely they are to correctly judge a model's recommendation as discriminatory.*

3.2.2. Strength of Evidence

In addition, we argue that for an explainable system, the strength of the evidence provided to support the explanation will influence the judgement the user reaches. Below, we highlight a couple of factors in this regard.

Miller and Gunasegaram (1990) argue that the temporal distance of events is an important moderating factor, specifically that people tend to "undo" more recent events. As it pertains to ethical decision making with AI, this would suggest that the user is unlikely to recognise the foil as valid if it is sufficiently temporally distant. Even if the case currently being assessed and the identified foil have identical properties, the user is likely to intuitively feel that due to the passage of time, changes to legislation, policies, and procedures, etc., treating the case as identical is not feasible. As a result, they may not dismiss evidence that points towards algorithmic discrimination, even if the user has some suspicion about the unethicality of previous decisions.

We posit that the use of probability in explanation, primarily when used to explain the causes of the recommendation, will increase the likelihood of correctly judging an algorithm's recommendation as discriminatory (see Josephson and Josephson 1996). The study by Eynon et al. (1997) would also appear to indicate that where ethical training is available, especially when it is tailored to the use of AI tools, users are likely to correctly judge a model's recommendation as discriminatory.

Proposition 7. *The more substantial the evidence presented to support an explanation, the more likely it is that the human agent will correctly judge the model as discriminatory as compared to weaker evidence.*

3.2.3. Social Pressures

Regarding the influence of social pressure on forming ethical judgements, Sonenshein argues that "organizations strongly influence how their members behave and what they believe". We posit that in terms of ethical decision making, a key influencing factor is the design of the AI tool and associated processes. Martin (2019) refers to these as "affordances—properties of technologies that make some actions easier than others". The higher the technological hurdle the user must clear, the less likely they are to adjudge the model's recommendation as discriminatory. For example, if the user has to perform more operations (e.g., navigate to different screens, click multiple buttons, etc.) in order to reject the AI tool's recommendation and provide a significant amount of mandatory justification (vis-à-vis accepting the recommendation), then the design of the tool or process is likely to influence their judgement. This concept has been acknowledged in "Values in Design" (ViD), which describes the field of research that investigates how "individually and organizationally held values become translated into design features" (Martin et al. 2019).

Proposition 8. *The more complicated the design of the tool and associated processes make it to judge the AI tool's recommendation as incorrect, the less likely it is that the human agent will do so.*

3.3. Explanation and Justification

Sonenshein (2007) posits that the judgement phase is followed by the explanation and justification phase, where the moral agent attempts to explain and justify their reaction to the constructed issue. We refer to this stage simply as the justification stage to avoid any confusion with the point(s) in the construction stage where the explainer provides an explanation for the recommendation. Sonenshein (2007) further argues that moral agents "employ the rules of rational analysis" to "bolster their confidence in the decision" as well as that of others. This reinforces the recommendation by Miller (2019) for the adoption of a conversational mode of explanation. The dialogue between the user and the explainer preserves an audit trail of the process by which the user constructed the issue and reached their judgement. Making this conversation readily available to the user for review also has the added advantage of highlighting inconsistencies in the issue construction process (e.g.,

implausible arguments). This highlights the requirement for social interaction between the explainee (i.e., the user) and the explainer.

Proposition 9. *The more conversational the dialogue between the explainer and the user, the better the quality of justification the human agent can provide for their judgement.*

4. Recommendations

Based on the preceding, we conclude with several non-exhaustive recommendations to assist users in making more ethically sound decisions when using AI tools. First, we recommend that the explanation provided by the explainer pre-empt the user's request for an explanation for abnormal events and make that available. Though Hilton (1990) recommends providing a contrastive explanation vis-à-vis a "typical" case, we suggest that the explainer should select an appropriate foil with identical properties and different recommendations and explain why different recommendations were made. The user should also have the ability to replace the system-selected foil with another they have selected and request a contrastive explanation for this. Though Miller (2019) suggests that an unprompted explanation could prove superfluous and distracting over time, we recommend that these explanations could be presented as "hints" where the details remain hidden but which can be readily accessed as required by the user. We argue that if a user correctly constructs and judges an issue as ethical early on, they are more likely to engage in moral behaviour and as such, investing in the design to highlight such issues is likely to drive desired behaviour.

Secondly, we propose the provision of explanations at different levels to facilitate ethical decision making. Apart from the explanations for each recommendation, the tool should be able to explain its model of self (see Section 3). Furthermore, the tool should also support the ability to request clarification on any section of the causal chain. The validation and verification of the tool should include expert users to test for the presence of algorithmic discrimination. Batarseh and Gonzalez (2015) detail a number of methods for accomplishing this, e.g., Context-Based Testing (CBT) and Turing Test approach.

Thirdly, based on the conversation model of explanation (see Hilton 1990), the explanation should be presented in a manner that follows the "basic rules of conversation". This would include only presenting information relevant to the user based on their mental model (see also Jaspars and Hilton 1988), keeping track of which information has already been shared (based on the premise that once something has been learned, it should not need to be explained again), and whether the user accepted it as credible or not, etc. It could also support presentation modes such as chatbots which facilitate conversational dialogue.

Fourthly, we recommend providing the capability for the user to collaborate with other users of the tool in the decision-making process. The user could share details of the case and the model's recommendation with one or more users and request their opinion(s). This would help to widen the initial user's frame of reference and could make them aware of more suitable foils. It would also assist in raising their experience level as the collective experience of all the collaborators will be utilised in the decision-making process.

Finally, we recommend that the process for rejecting the AI model's recommendation and highlighting the potential existence of ethical issues should be as streamlined as possible and should not be more complicated than the process for accepting the model's recommendation. This will increase the likelihood that the user will follow through on any moral intent they had previously established.

5. Limitation and Future Research

In this paper, we have synthesised the literature on ethical decision making and explainable AI and proposed several testable belief statements. We believe the main limitation of this study is the lack of empirical evidence to support the belief statements. However, we expect that these will be tested and empirically validated in future studies utilising a variety of appropriate methodologies. For example, there exist opportunities to

utilise technology such as functional Magnetic Resonance Imaging (fMRI) to monitor the brain activity of users as they make decisions using AI tools, undertake a phenomenological study to gather rich data among real-life practitioners, amongst others. In future work, we intend to attempt to tackle a number of these opportunities.

Secondly, rather than a systematic search of the ethical decision making and Explainable AI literature base, we undertook a top-down search starting from key papers in both fields and exploring the linkages from these. We selected this approach as this search strategy is recommended as the most effective way of reviewing and synthesizing research fields which are different. However, we acknowledge that the risk exists that we may have missed relevant studies in both fields. We have mitigated this risk by obtaining feedback from experts who have provided valuable feedback to ensure completeness of the review.

Author Contributions: Conceptualisation, N.O.; methodology, N.O.; software, N.O.; validation, A.B. and T.C.; formal analysis, N.O.; investigation, N.O.; resources, N.O., A.B., and T.C.; data curation, N.O.; writing—original draft preparation, N.O.; writing—review and editing, A.B. and T.C.; visualisation, N.O.; supervision, A.B. and T.C.; project administration, N.O. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare no conflict of interest.

References

Association for the Advancement of Artificial Intelligence. 2019. *AAAI Code of Professional Ethics and Conduct*. Available online: https://aaai.org/Conferences/code-of-ethics-and-conduct.php (accessed on 11 April 2020).

Abdul, Ashraf, Jo Vermeulen, Danding Wang, Brian Y. Lim, and Mohan Kankanhalli. 2018. Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda. Paper presented at 2018 CHI Conference on Human Factors in Computing Systems, Montreal, QC, Canada, April 21–26; pp. 1–18.

Academy of Medical Science. 2017. *House of Commons Briefing Paper 351, 2017–2019, ALG0055*. London: UK Parliament.

Adadi, Amina, and Mohammed Berrada. 2018. Peeking inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access* 6: 52138–60. [CrossRef]

Amoore, Louise. 2017. Algorithms in Decision-Making, 2018, May 23. In *House of Commons Briefing Paper 351, 2017–2019, Q4*. London: UK Parliament.

Batarseh, Feras A., and Avelino J. Gonzalez. 2015. Validation of Knowledge-Based Systems: A Reassessment of the Field. *Artificial Intelligence Review* 43: 485–500. [CrossRef]

Beauchamp, Tom L., Norman E. Bowie, and Denis Gordon Arnold, eds. 2004. *Ethical Theory and Business*. London: Pearson Education.

Bell, Emily. 2016. Controlling the Unaccountable Algorithm. *BBC*. Available online: https://www.bbc.co.uk/sounds/play/b085wj18 (accessed on 23 March 2020).

Bowie, Norman E. 2017. *Business Ethics: A Kantian Perspective*. Cambridge: Cambridge University Press.

Cavanagh, Gerald F., Dennis J. Moberg, and Manuel Velasquez. 1981. The Ethics of Organisational Politics. *Academy of Management Review* 6: 363–74.

De Graaf, Maartje M. A., and Bertram F. Malle. 2017. How People Explain Action (and Autonomous Intelligent Systems Should Too). Paper presented at 2017 AAAI Fall Symposium Series, Arlington, TX, USA, November 9–11.

Dodd, David H., and Jeffrey M. Bradshaw. 1980. Leading Questions and Memory: Pragmatic Constraints. *Journal of Verbal Learning and Verbal Behavior* 19: 695–704. [CrossRef]

Eynon, Gail, Nancy Thorley Hills, and Kevin T. Stevens. 1997. Factors That Influence the Moral Reasoning Abilities of Accountants: Implications for Universities and the Profession. *Journal of Business Ethics* 16: 1297–309. [CrossRef]

Ferrell, Odies C., and Larry G. Gresham. 1985. A Contingency Framework for Understanding Ethical Decision Making in Marketing. *Journal of Marketing* 49: 87–96. [CrossRef]

Financial Service Consumer Panel. 2017. *House of Commons Briefing Paper 351, 2017–2019*. London: UK Parliament.

Fiske, Susan T., and Shelley E. Taylor. 1991. *Social Cognition*. New York: Mcgraw-Hill Book Company.

Gilbert, Daniel T., and Patrick S. Malone. 1995. The Correspondence Bias. *Psychological Bulletin* 117: 21. [CrossRef]

Gilpin, Leilani H., David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining Explanations: An Overview of Interpretability of Machine Learning. Paper presented at 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), Turin, Italy, October 1–3; pp. 80–89.

Harman, Gilbert H. 1965. The Inference to the Best Explanation. *The Philosophical Review* 74: 88–95. [CrossRef]

Hassani, Hossein, Emmanuel Sirimal Silva, Stephane Unger, Maedeh TajMazinani, and Stephen Mac Feely. 2020. Artificial Intelligence (AI) or Intelligence Augmentation (IA): What Is the Future? *AI* 1: 143–55. [CrossRef]

Hayes, Bradley, and Julie A. Shah. 2017. Improving Robot Controller Transparency through Autonomous Policy Explanation. Paper presented at 2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI), Vienna, Austria, March 6–9; pp. 303–12.

Hegarty, W. Harvey, and Henry P. Sims. 1978. Some Determinants of Unethical Decision Behavior: An Experiment. *Journal of Applied Psychology* 63: 451. [CrossRef]

Hilton, Denis J. 1990. Conversational processes and causal explanation. *Psychological Bulletin* 107: 65. [CrossRef]

Hilton, Denis J. 1996. Mental Models and Causal Explanation: Judgements of Probable Cause and Explanatory Relevance. *Thinking & Reasoning* 2: 273–308.

Hilton, Denis J., and Ben R. Slugoski. 1986. Knowledge-Based Causal Attribution: The Abnormal Conditions Focus Model. *Psychological Review* 93: 75. [CrossRef]

Hunt, Shelby D., and Scott Vitell. 1986. A General Theory of Marketing Ethics. *Journal of Macromarketing* 6: 5–16. [CrossRef]

IEEE. 2016. IEEE Standards Association Introduces Global Initiative for Ethical Considerations in the Design of Autonomous Systems. Available online: https://standards.ieee.org/news/2016/ieee_autonomous_systems.html (accessed on 25 March 2020).

Jaspars, Jospeh M., and Dennis J. Hilton. 1988. Mental Models of Causal Reasoning. In *The Social Psychology of Knowledge*. Editions de la Maison des Sciences de l'Homme. Edited by D. Bar-Tal and A. W. Kruglanski. Cambridge: Cambridge University Press, pp. 335–58.

Johnson, Deborah G. 2015. Technology with No Human Responsibility? *Journal of Business Ethics* 127: 707–15. [CrossRef]

Jones, Thomas M. 1991. Ethical Decision Making by Individuals in Organisations: An Issue-Contingent Model. *Academy of Management Review* 16: 366–95. [CrossRef]

Josephson, John R., and Susan G. Josephson, eds. 1996. *Abductive Inference: Computation, Philosophy, Technology*. Cambridge: Cambridge University Press.

Kauffmann, Jacob, Klaus-Robert Müller, and Grégoire Montavon. 2020. Towards Explaining Anomalies: A Deep Taylor Decomposition of One-Class Models. *Pattern Recognition* 101: 107198. [CrossRef]

Kelman, Herbert C., and V. Lee Hamilton. 1989. *Crimes of Obedience: Toward a Social Psychology of Authority and Responsibility*. New Haven and London: Yale University Press.

Klimov, Pavel. 2017. Algorithms in Decision-Making, 2018. May 23. In *House of Commons Briefing Paper 351, 2017–2019, Q83*. London: UK Parliament.

Kulesza, Todd, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. 2013. Too Much, Too Little, or Just Right? Ways Explanations Impact End Users' Mental Models. Paper presented at 2013 IEEE Symposium on Visual Languages and Human Centric Computing, San Jose, CA, USA, September 15–19; pp. 3–10.

Li, Xiao-Hui, Caleb Chen Cao, Yuhan Shi, Wei Bai, Han Gao, Luyu Qiu, Cong Wang, Yuanyuan Gao, Shenjia Zhang, Xun Xue, and et al. 2020. A Survey of Data-Driven and Knowledge-Aware Explainable Ai. *IEEE Transactions on Knowledge and Data Engineering*. [CrossRef]

Martin, Kirsten. 2019. Ethical Implications and Accountability of Algorithms. *Journal of Business Ethics* 160: 835–50. [CrossRef]

Martin, Kirsten, Katie Shilton, and Jeffery Smith. 2019. Business and the Ethical Implications of Technology. *Introduction to the Symposium*, 1–11.

Maybin, S. 2016. How Maths Can Get You Locked Up. Available online: https://www.bbc.co.uk/news/magazine-37658374 (accessed on 2 April 2020).

Messick, David M., and Max H. Bazerman. 1996. Ethical Leadership and the Psychology of Decision Making. *MIT Sloan Management Review* 37: 9.

Miller, Tim. 2019. Explanation in Artificial Intelligence: Insights from the Social Sciences. *Artificial Intelligence* 267: 1–38. [CrossRef]

Miller, Dale T., and Saku Gunasegaram. 1990. Temporal Order and the Perceived Mutability of Events: Implications for Blame Assignment. *Journal of Personality and Social Psychology* 59: 1111. [CrossRef]

Murgia, M. 2019. Algorithms Drive Online Discrimination, Academic Warns. *Financial Times*. Available online: https://www.ft.com/content/bc959e8c-1b67-11ea-97df-cc63de1d73f4 (accessed on 10 March 2020).

Oxford Internet Institute. 2017. Algorithms in Decision-Making, 2018, May 23. In *House of Commons Briefing Paper 351, 2017–2019, ALG0031*. London: UK Parliament.

Paradice, David B., and Roy M. Dejoie. 1991. The Ethical Decision-Making Processes of Information Systems Workers. *Journal of Business Ethics* 10: 1–21. [CrossRef]

Park, Dong Huk, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. 2018. Multimodal Explanations: Justifying Decisions and Pointing to the Evidence. Paper presented at IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, June 18–22; pp. 8779–88.

Peachey, Kevin. 2019. Sexist and Biased? *How Credit Firms Make Decisions*. Available online: https://www.bbc.co.uk/news/business-50432634 (accessed on 15 March 2020).

Peirce, Charles Sanders. 1997. *Pragmatism as a Principle and Method of Right Thinking: The 1903 Harvard Lectures on Pragmatism*. Albany: SUNY Press.

Rest, James R. 1986. *Moral Development: Advances in Research and Theory*. New York: Praeger.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. 2016. Why Should I Trust You?" Explaining the Predictions of Any Classifier. Paper presented at 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13–17; pp. 1135–44.

Schweitzer, Maurice E., Lisa Ordóñez, and Bambi Douma. 2004. Goal Setting as a Motivator of Unethical Behavior. *Academy of Management Journal* 47: 422–32.

Science and Technology Committee. 2018. Algorithms in Decision-Making. 2018, May 23. In *House of Commons Briefing Paper 351, 2017–2019*. London: UK Parliament.

Smith-Crowe, Kristin. 2004. An Interactionist Perspective on Ethical Decision-Making: Integrative Complexity and the Case of Worker Safety. Ph.D. dissertation, Tulane University, New Orleans, LA, USA.

Sonenshein, Scott. 2007. The Role of Construction, Intuition, and Justification in Responding to Ethical Issues at Work: The Sense-Making-Intuition Model. *Academy of Management Review* 32: 1022–40. [CrossRef]

Strudler, Alan, and Danielle E. Warren. 2001. Authority, heuristics, and the structure of excuses. In *Next Phase of Business Ethics: Integrating Psychology and Ethics*. Greenwich: JAI Press, pp. 355–75.

Tenbrunsel, Ann E. 1998. Misrepresentation and Expectations of Misrepresentation in an Ethical Dilemma: The Role of Incentives and Temptation. *Academy of Management Journal* 41: 330–39.

Tenbrunsel, Ann E., and Kristin Smith-Crowe. 2008. 13 Ethical Decision Making: Where We've Been and Where We're Going. *The Academy of Management Annals* 2: 545–607. [CrossRef]

Tenbrunsel, Ann E., Kristin Smith-Crowe, and Elizabeth E. Umphress. 2003. Building Houses on Rocks: The Role of the Ethical Infrastructure in Organisations. *Social Justice Research* 16: 285–307. [CrossRef]

Trevino, Linda Klebe. 1986. Ethical Decision Making in Organisations: A Person-Situation Interactionist Model. *Academy of Management Review* 11: 601–17. [CrossRef]

Treviño, Linda K., Gary R. Weaver, and Scott J. Reynolds. 2006. Behavioral ethics in organizations: A review. *Journal of Management* 32: 951–90. [CrossRef]

Tversky, Amos, and Daniel Kahneman. 1983. Extensional versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgment. *Psychological Review* 90: 293. [CrossRef]

Van Bouwel, Jeroen, and Erik Weber. 2002. Remote Causes, Bad Explanations? *Journal for the Theory of Social Behaviour* 32: 437–49. [CrossRef]

Weick, Karl E. 1993. The Collapse of Sensemaking: The Mann Gulch Disaster. *Administrative Science Quarterly* 38. [CrossRef]

Weller, Adrian. 2017. Algorithms in Decision-Making. 2018, May 23. In *House of Commons Briefing Paper 351, 2017–2019, Q30*. London: UK Parliament.

Xu, Feiyu, Hans Uszkoreit, Yangzhou Du, Wei Fan, Dongyan Zhao, and Jun Zhu. 2019. Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges. Paper presented at CCF International Conference on Natural Language Processing and Chinese Computing, Dunhuang, China, October 9–14; Berlin: Springer, pp. 563–74.